DOCUMENT RESUME

ED 169 097 TM 008 493

AUTHOR	Steinheiser, Frederick H., Jr.; And Others
TITLE	Criterion-Referenced Testing: A Critical Analysis of Selected Models. Technical Paper 306. Final Report
INSTITUTION	Army Research Inst. for the Behavioral and Social Sciences, Alexandria, Va.
SPONS AGENCY	Office of the Deputy Chief of Staff for Personnel (Army), Washington, D.C.
REPORT NO	ARI-TP-306
PUB DATE	Aug 78
CONTRACT	DAHC-19-75-M-0003
NOTE	67p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	Comparative Statistics; *Criterion Referenced Tests; *Cutting Scores; Decision Making; *Mastery Tests; *Mathematical Models; Minimum Competency Testing; *Predictive Measurement; *Statistical Analysis; Test Interpretation; Validity
IDENTIFIERS	Test Length

ABSTRACT

Alternative mathematical models for scoring and decision making with criterion referenced tests are described, especially as they concern appropriate test length and methods of establishing statistically valid cutting scores. Several of these approaches are reviewed and compared on formal-analytic and empirical grounds: (1) Block's approach to setting mastery standards, involving student performance and academic requirements; (2) Crehan's classification, comparing scores of students who have and who have not completed training; (3) the probabilistic models of Macready, Dayton, and Emrick, which assume an equal proportion of masters and nonmasters; (4) the binomial distribution model, which allows for partial acquisition of proficiency; (5) the Bayesian model, which considers prior experience; (6) Rasch's one-parameter logistic model, which yields person-free test calibrations and item-free person measurements; and (7) the regression approach of classical test theory, which enables the estimation of true scores to be made from observed scores. Examples of these approaches are given, as well as their advantages, disadvantages, and ambiguities. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED169097

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

AD

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Technical Paper 306

CRITERION-REFERENCED TESTING: A CRITICAL ANALYSIS OF SELECTED MODELS

Frederick H. Steinheiser, Jr., Kenneth I. Epstein, and Angelo Mirabella

and

George B. Macready, University of Maryland

UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

TM008 493

U. S. Army

Research Institute for the Behavioral and Social Sciences

August 1978

Approved for public release; distribution unlimited.

U. S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES

A Field Operating Agency under the Jurisdiction of the Deputy Chief of Staff for Personnel

JOSEPH ZEIDNER
Technical Director (Designate)

WILLIAM L. HAUSER
Colonel, US Army
Commander

Research accomplished
under contract to the Department of the Army

University of Maryland

NOTICES

DISTRIBUTION: Primary distribution of this report has been made by ARI. Please address correspondence concerning distribution of reports to: U. S. Army Research Institute for the Behavioral and Social Sciences, ATTN: PERI-P, 5001 Eisenhower Avenue, Alexandria, Virginia 22333.

FINAL DISPOSITION: This report may be destroyed when it is no longer needed. Please do not return it to the U. S. Army Research Institute for the Behavioral and Social Sciences.

NOTE: The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Technical Paper 306	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) CRITERION-REFERENCED TESTING: A CRITICAL ANALYSIS OF SELECTED MODELS		5. TYPE OF REPORT & PERIOD COVERED Final Report 6. PERFORMING ORG. REPORT NUMBER --
7. AUTHOR(s) Frederick H. Steinheiser, Jr., Kenneth I. Epstein, and Angelo Mirabella (ARI); George B. Macready (University of Maryland)		8. CONTRACT OR GRANT NUMBER(s) DAHC19-75-M-0003
9. PERFORMING ORGANIZATION NAME AND ADDRESS University of Maryland, College Park, Maryland ARI, 5001 Eisenhower Ave., Alex., VA		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 2062722A764
11. CONTROLLING OFFICE NAME AND ADDRESS Deputy Chief of Staff for Personnel Washington, D.C. 20310		12. REPORT DATE August 1978 13. NUMBER OF PAGES 56
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) --		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE --

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

--

18. SUPPLEMENTARY NOTES

--

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Criterion-referenced testing
Decisionmaking
Bayesian statistics
Rasch model
Classification error

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Several mathematical models for use in criterion-referenced testing are reviewed and compared. The models are evaluated on both formal-analytic and empirical grounds. Predictive models include probabilistic formulations, a binomial model, and a Bayesian model. Descriptive methods include a categorization scheme, a one-parameter logistic model, and linear regression. An empirical method for relating mastery criteria to derived educational outcomes is also included. Problems inherent in each model or class of models are described. Such problems

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20.

include tenability of assumptions, ease of application, assessment of item characteristics, and assessment of the model's fit to data. Each method/model appears to be appropriate for specific types of testing situations, although further development will depend upon computer simulation and empirical research.

Technical Paper 306

CRITERION-REFERENCED TESTING: A CRITICAL ANALYSIS OF SELECTED MODELS

Frederick H. Steinheiser, Jr., Kenneth I. Epstein and Angelo Mirabella

and

George B. Macready, University of Maryland

UNIT TRAINING AND EVALUATION SYSTEMS TECHNICAL AREA

Submitted as complete and
technically accurate, by:
Frank J. Harris
Technical Area Chief

Approved By:

A.H. Birnbaum, Acting Director
ORGANIZATIONS AND SYSTEMS
RESEARCH LABORATORY

Joseph Zeidner
TECHNICAL DIRECTOR (DESIGNATE)

U.S. ARMY RESEARCH INSTITUTE FOR THE BEHAVIORAL AND SOCIAL SCIENCES
5001 Eisenhower Avenue, Alexandria, Virginia 22333

Office, Deputy Chief of Staff for Personnel
Department of the Army

August 1978

Army Project Number
2Q62722A769

Unit Training Standards
and Evaluation

Approved for public release; distribution unlimited.

ARI Research Reports and Technical Papers are intended for sponsors of R&D tasks and other research and military agencies. Any findings ready for implementation at the time of publication are presented in the latter part of the Brief. Upon completion of a major phase of the task, formal recommendations for official action normally are conveyed to appropriate military agencies by briefing or Disposition Form.

The research presented in this report was conducted under Project METTEST (Methodological Issues in Criterion-Referenced Testing), in the Unit Training and Evaluation Systems (UTES) Technical Area of ARI under Army RDTE Project 2Q62722A764. The goal of Project METTEST is to provide quantitative methods for evaluating unit proficiency. The means for achieving this goal include basic research in test construction methodology, measurement and scaling models, and decisionmaking implications of test score interpretation.

Related, ongoing programs within the UTES Technical Area include evaluation of small combat units under simulated battlefield conditions (REALTRAIN, ARTEP), qualification of tank crews and platoon gunnery (IDOC), and improvement of the reliability of ARTEP evaluation.

Anticipated future research under Project METTEST includes the development of a computer model for performance evaluation, and development of measurement, scaling, scoring, decisionmaking, and quality control models for use in performance evaluations when criterion-referenced testing procedures are employed.

ARI research in this area is conducted as an in-house research effort augmented by contracts with organizations selected as having unique capabilities and facilities for research in a specific area. The present study was conducted in collaboration with personnel of the University of Maryland under Contract No. DAHC19-75-M-0003.

JOSEPH ZEIDNER
Technical Director (Designate)

CRITERION-REFERENCED TESTING: A CRITICAL ANALYSIS OF SELECTED MODELS

BRIEF

Requirement:

To develop a theoretical base for research and eventual application of methods for assigning pass-fail scores in personnel and unit evaluation using the criterion-referenced testing approach.

Procedure:

Relevant literature for each of five approaches to criterion-referenced testing was reviewed. The approaches were compared on the basis of the following: assumptions and rationale, the interactive effects of test length and passing criteria on classification accuracy, and areas of applicability. A computational example was prepared for each model, and strengths and weaknesses were also evaluated.

Findings:

Four of the five models were able to specify an "optimal" test length and cutoff score, although they differed as to the required parameter estimates from the test developer. For example, expert "prior" information can be used to reduce test length. Each of the models also provides an estimate for misclassifications, or Type I and Type II errors. The models are neither redundant nor interchangeable. No "best" method was identified. Rather, the selection of a model depends upon the particular measurement requirements and constraints as identified by the test developer.

Utilization of findings:

This research provides qualitative and quantitative guidelines for developers of criterion-referenced tests. The models have been applied to analyze data from the handgun qualification course at the U.S. Army Military Police School. Application of the models has also been addressed to revision of Table VIII tank gunnery.

CRITERION-REFERENCED TESTING: A CRITICAL ANALYSIS OF SELECTED MODELS

CONTENTS

	Page
INTRODUCTION	1
Theoretical Problems for CRT Models	5
Overview of Selected CRT Models	6
REVIEW OF MODELS	6
Block	6
Crehan	7
Macready and Dayton - Emrick	8
Binomial Model	15
Bayesian Model	19
Rasch's One-Parameter Logistic Model	29
Regression Theory	37
SUMMARY AND CONCLUSIONS	40
Nature of Performance Acquisition	40
Measurement Error	40
Classification Error	41
Test Length	42
Conceptualization of Mastery	43
Characteristics of Items	43
REFERENCES	45
APPENDIXES	
A. A Generalization of the Emrick Model for the Case of Unequal Proportions of Masters and Nonmasters Kenneth I. Epstein	49
B. Critique of the Simplifying Assumptions in Using Regression Models for Estimating True Scores from Observed Scores James McBride	53
DISTRIBUTION	55

Page

LIST OF TABLES

Table 1. Summary comparison of some methods and models used in criterion-referenced testing 3

2. Example data matrixes for the Crehan procedure 9

3. Probability of observing response patterns under the Macready and Dayton models, assuming $P(M) = P(\bar{M}) = .5$ 12

4. Probability of misclassification as a function of cutting score under the Macready and Dayton models, assuming $P(M) = P(\bar{M}) = .5$ 14

5. Changes in posterior probability of mastery as a function of changes in prior probability of mastery 22

6. Cumulative estimation of prior probabilities for various assumed proficiencies 26

7. Point values for prior proficiency distribution 28

A-1. Hypothetical response data for masters and nonmasters 49

A-2. Measurement errors and mastery state for hypothetical data 50

A-3. Table of proportions for observed responses and mastery state in terms of α, β, $P(M)$ and $P(\bar{M})$. 51

LIST OF FIGURES

Figure 1. "Person-bound" test calibration using non-overlapping groups of masters and nonmasters (adapted from Wright, 1967) 32

2. "Person-free" test calibration using the Rasch logistic model ("o" = master, "o" = nonmaster; adapted from Wright, 1967) 33

3. Regression of true score on observed score for a five-item test 38

CRITERION-REFERENCED TESTING: A CRITICAL ANALYSIS OF SELECTED MODELS

INTRODUCTION

Scoring and decisionmaking models for criterion-referenced testing deal with two questions of practical and theoretical importance: (1) how much test information should be collected to provide a basis for confident decisions about the mastery or nonmastery of trained skills; and (2) what are the methods of establishing statistically valid standards of achievement. Criterion-referenced testing (CRT) requires that the data provide information about performance capabilities measured against some external criterion (Glaser & Nitko, 1971; Carver, 1974). Such criteria are properly derived from an analysis of the requirements for performing specific tasks successfully.

Measurement of mastery implies that CRT's should represent the skills to be measured with high fidelity. However, serious constraints are imposed by requiring high fidelity: (1) the time needed to administer the test may be more than is readily available; (2) the number of examiners needed to administer the test and collect data may be excessive; (3) the expenditure of materials used in testing may be prohibitively high; and (4) the appropriate testing materials or apparatus may not be available for a long enough time. These constraints place a premium upon limiting test data to the minimum amount sufficient for the desired quality of decisionmaking. Statistical models offer one means of accomplishing this goal.

Two problems arise in establishing achievement standards on CRT's. The first is related to the congruence between CRT performance and real-world requirements. The second is related to the statistical inferences applied to observed CRT scores.

Before any statistical model can be used in a CRT situation, the requirements for mastery over the domain in general must be specified. The requirements usually describe the capabilities of persons who can successfully perform the tasks included in the domain. Glaser and Klaus (1963) suggest that "proficiency standards can be established at any value between the point where the system will not perform at all and the point where any further contribution from the human component will not yield any increase in system performance (p. 424)."

These system requirements may include the human performance components of industrial-vocational tasks, minimal competencies in an educational system, or basic literacy skills. System requirements may also reflect manpower needs, the criticality of the task, or the consequences of poor performance. Such idealized standards must then be converted to standards on a particular CRT. The conversion process

involves issues of test validity which are beyond the scope of this paper. Meskauskas (1976) discusses several methods that have been used to bridge the gap between operational tests and real-world requirements.

If the CRT includes the entire full fidelity task, such as disassembling and cleaning a particular piece of machinery, then setting mastery standards is relatively clear and unambiguous. However, if the CRT includes only a sample of the full fidelity task, or if fidelity is decreased for practical purposes, then mastery standards for the CRT are not clearcut. Heretofore, the use of arbitrary cutoff scores has kept this problem at a manageable level. For example, objectives often include a statement of standards requiring a certain minimum percent correct for attainment of mastery status. Two criticisms can be directed at this concept of mastery.

First, any percentage correct is a relative standard. The definition of mastery has been shown (Millman, 1972; Novick & Lewis, 1974; Epstein & Steinheiser, 1975) to be a function both of the percentage correct and of the number of trials or items that comprise the test. A more comprehensive definition could be based either upon (1) an idealization, such as the proportion of correct answers of all possible test items, or (2) the position on an underlying continuum of ability hypothesized to score an examinee on a given test. By stating standards in terms of such an idealization or ability continuum, it is possible to explicitly define mastery cutoff scores for any test length.

The second criticism refers to the level of ability required for mastery. For example, why should one standard (such as 80% correct) be set rather than another (such as 70% or 90%)? Perhaps this question could be answered by empirical studies showing the relationship between CRT scores and the transfer or retention of training. The required level of mastery could also be determined by system requirements, criticality, and similar factors.

Each of the models discussed in this paper, with the exception of Block's (1972) approach to setting standards empirically, assumes that a well-defined universe of items exists or can be generated. The authors also assume that the role of the statistical model is to describe accurately an examinee with respect to that universe. The validity of the generalization from the universe of items to the real world is not investigated. The models further assume that a mastery standard relative to the entire universe can be established. Given these assumptions, the problem is how to interpret the observations. The following section discusses theoretical issues which may produce possible solutions. Table 1 then introduces and summarizes the specific models.

The problem of setting standards arises because it is often impractical to insist upon complete mastery of a task, or even to require a very high percentage of correct answers to the items comprising a CRT. Furthermore, it is often impossible to list all of the potential items

Table 1

Summary Comparison of Some Methods and Models Used in Criterion-Referenced Testing

Model	Nature of performance acquisition	Theoretical observed score: x = score, n = # items, A = true ability	True score distribution	Cutoff score specification
Block	Undefined	Undefined	Undefined	Empirical, based upon external criterion
Crehan	Undefined	Pre-instr: x = 0 Post-instr: x = n	Dichotomous, based on pre-post instruction	Empirical; pre-post instruction classification
Emrick	All-or-none	Nonmaster: x = 0 Master: x = n	Dichotomous, master or nonmaster	Choose score that best dichotomizes observed score distribution, assuming guessing and forgetting errors.
Dayton & Macready	All-or-none	Nonmaster: x = 0 Master: x = n	Dichotomous, master or nonmaster	Choose score that best dichotomizes observed score distribution, assuming guessing and forgetting errors.
Kriewall-Millman (Binomial)	Continuous	$p(x\|A) = \binom{n}{x} A^x (1 - A)^{n - x}$	Undefined	Choose score such that the sum of probability of achieving at least that score for nonmasters, and not achieving that score for masters is minimized.
Novick et al. (Bayesian)	Continuous	$p(x\|A) = \binom{n}{x} A^x (1 - A)^{n - x}$	Beta-binomial	Calculate posterior probability that observed score exceeds the standard.

Table 1 (continued)

Summary Comparison of Some Methods and Models Used in Criterion-Referenced Testing

Model	Nature of performance acquisition	Theoretical observed score: x = score, n = # items, A = true ability	True score distribution	Cutoff score specification
Rasch (logistic)	Continuous	$p(x\|A) = \prod_{i=1}^{n} \frac{e^{(b_i - A)}}{1 + e^{(b_i - A)}}$ b_i = item difficulty	Normal	Choose minimum Rasch ability estimate. Calculate the ability estimate from observed score.
Classical regression	Continuous	$x = A - e$, where e = error of measurement	Normal	Choose minimum "true" score criterion. Calculate estimated true score from observed score.

of a given task domain. For example, an indefinitely large number of multiplication items could comprise an item universe from which a sample of items are selected. An arbitrary standard would determine that the examinee answering a specified number (or percentage) of the sample correctly will be classified as a "master" of multiplication. The main purpose of the present paper is to evaluate several mathematical models that claim to reduce the arbitrariness in setting criteria for mastery on tests representing a sample of the test-item universe. The motivation for developing models by which criteria for mastery can be derived formally arises from the goal of trying to minimize misclassifications (i.e., designating a "true master" as a "nonmaster" or vice versa). The more complex the skills assessed by the CRT, the smaller the sample of items, and the more varied the type of performance included in the universe, the greater the danger of misclassification.

Theoretical Problems for CRT Models

Nature of Performance Acquisition. Is the attainment of mastery an "all-or-none" occurrence, or is there a continuum of varying degrees of skill acquisition? The widely accepted dichotomy of master vs. nonmaster may be overly simplistic. The alternative is a continuum of varying degrees of mastery. Both dichotomous and continuous CRT models are available in the literature.

Measurement Error. One type of error, similar to the classical psychometric notion of measurement error, refers to random inappropriate responses due to temporary environmental distractions, lucky guesses, lapses in attention, etc. The magnitude of such error can be estimated and included in the estimation of actual ability and in the determination of test standards and lengths.

A second type, "classification" error, refers to the (usually) dichotomous classification of an examinee as a master or nonmaster. Its magnitude and direction are primarily a function of how a cutoff score is chosen. Classification error will tend to increase as the accuracy in estimating actual ability decreases, but a mathematically defined relationship between measurement error and classification error has not been derived (Guilford, 1956, pp. 380-384).

Test Length to Distinguish Masters from Nonmasters. One technique to improve ability estimation and reduce the chance for misclassification is to increase the number of test items. In some situations this may be possible simply by repeating items until the desired level of precision is attained. However, in most cases, test length cannot be indefinitely increased. Therefore, a statistical model that provides increased information per item is highly desirable. Generally, a CRT model should provide sufficient information to decisionmakers so that they will know the risks of committing false positive and false negative errors before the test is conducted.

Overview of Selected CRT Models

The CRT models discussed in this paper were chosen to try to illustrate the diversity in approaches to the problems outlined in the preceding section. Methods developed by Crehan (1974) and Block (1972) are basically empirical in that cutoff scores are based upon empirically derived requirements. Models derived by Emrick (1971) and by Macready and Dayton (1976) assume a dichotomous definition of mastery and analytically describe procedures for establishing cutoff scores. Kriewall (1969) and Millman (1972, 1974) assume that responses to test items and examinee ability can be described by the family of binomial distributions. Their basic models can be extended by applying the theory of binomial error models (Lord & Novick, 1968). Novick and Lewis (1974) discuss the application of a Bayesian approach to CRT issues. A one-parameter logistic model (Rasch, 1960; Wright, 1967) provides a practical example of how latent trait theory may be applied to CRT data analysis. Finally, an approach for CRT data analysis derived from classical regression theory is discussed. Each model is examined in terms of rationale and assumptions, empirical support and applications, illustrative examples of the type of input required and output provided, and critical evaluation.

REVIEW OF MODELS

Block

Block's (1972) research provides an experimental approach to setting mastery standards. He studied the relationship between the level of performance required on each unit of a three-unit instructional sequence and five cognitive and affective outcome variables. The rationale for this study was the intuitive notion that maximum performance on an external measure of achievement would be observed in students having the most stringent passing requirements in the instruction. A second question concerned the relationship between scores on an affective measure of interest and attitude and passing requirements in instruction.

Block's experiment included four treatment groups that differed from one instructional unit to the next with respect to the standard required for advancement. If the student did not meet the standard (65%, 75%, 85%, or 95% of the items correct on a 20-item test), remedial instruction was provided. Students in a control group proceeded from one unit to the next with no remediation, regardless of their test score. Five outcome variables were defined: achievement, learning rate, transfer, interest, and attitude.

Transfer was measured by a 10-item test which required the use of the learned skills to solve a novel set of problems. It was given both as a pretest and after instruction. Interest and attitude were measured using a 24-item questionnaire.

Most of the results supported the intuitive hypothesis. The control group did consistently worse on achievement, transfer, and retention, than any of the experimental groups, and the learning curves suggested that high standards early in an instructional sequence may produce increased efficiency later in the sequence. However, several interesting exceptions to the intuitive expectations suggest that higher standards are not always better standards. For example, the 85% and 95% groups did not differ from one another on retention or achievement measures, although they both differed from the control group. Only the 85% group produced sustained high levels of interest and attitude.

Block's research suggests that a unitary definition of an "optimum" CRT cutting score may be questionable. If uniformly high achievement and transfer are required at the possible expense of positive interest and attitude, it may be that the highest mastery standard should be used. However, if some "mix" of cognitive and affective outcomes is desired, then a lower standard seems appropriate.

Similar studies could be conducted on a wide range of instructional programs for a wide variety of outcomes. The results could lead to usable and meaningful guidelines for setting cutting scores to optimize a number of instructional outcomes. Because the results may not be generalizable across content areas and instructional programs, such an optimization strategy would require costly and extensive research. This empirical verification of a decisionmaking strategy for finding optimal mixes of cognitive and affective outcomes does not mathematically model any of the problems outlined in the previous section of this paper. A truly complete scoring and decisionmaking CRT model would take into account both the psychological variables that characterize optimum learning and the constraints imposed by test length, cutting scores, and misclassification rates.

Crehan

A method used by Crehan (1974) also relies heavily on a training context for its interpretation. The method's rationale for specifying cutting scores is based upon the comparison of the test scores of students who have completed training with the test scores of those who have not yet received training. This method provides a means of assessing the proportion of misclassified students within each group when various cutting scores are used.

Correct classification occurs when posttraining students pass the test and students with no training fail the test. Using a 2 x 2 matrix of pass-fail and training-no training for each cutting score, the proportion of correct classifications P_c can be obtained as follows:

P_c = [number who had training and passed + number who had no training and failed] ÷ sum of all four entries in the matrix.

A cutting score is found by choosing the score that maximizes the proportion of correct classifications.

For example, assume that the distribution of scores on a five-item CRT for an untrained group and a group that has completed training is as follows:

Number Correct	No Training	Completed Training
0	10	0
1	5	0
2	4	1
3	0	5
4	1	10
5	0	4

A series of fourfold tables in Table 2 displays the relationships between cutting score, pass-fail decisions, and the amount of training. P_C, the proportion of correct classifications, is calculated for each fourfold table. The highest value of P_C in this example is found when three correct responses are used as the cutting score. Therefore, for this training program, a cutting score of 3 would be recommended as the optimal cutting score.

The major strength of this procedure is that it provides an estimate of the optimal cutting score for differentiating between trained and untrained groups while remaining relatively simple to implement. However, these two groups do not necessarily correspond to the categories of "masters" and "nonmasters" in terms of the ability of group members to complete an objective. Instead, one might expect the posttraining group to perform less well than a group consisting entirely of examinees who have mastered the objective, and the pretraining group to perform somewhat better than a group of examinees, none of whom has mastered the objective.

The simplicity of Crehan's procedure is partially offset by a number of weaknesses, including the following: (1) lack of a procedure for estimating the minimum item sample size necessary to keep the probability of misclassification at or below some specified level; and (2) lack of statistical criteria for differentiating between P_C's which "seem" to be similar (or different).

Macready and Dayton - Emrick

Assumptions and Rationale. Two related probabilistic models that provide probability estimates of the 2^n possible response patterns on a dichotomously scored, n-item test are discussed in this section (Emrick, 1971; Dayton & Macready, 1976; and Macready & Dayton, 1975). Both models assume that all examinees belong to one of two possible

Table 2

Example Data Matrices for the Crehan Procedure

Cutting score		Training experience: No training	Training experience: Completed training
0	Pass	20	20
	Fail	0	0
	Pc = 20/40 = .5		
1	Pass	10	20
	Fail	10	0
	Pc = 30/40 = .75		
2	Pass	5	20
	Fail	15	0
	Pc = 35/40 = .875		
3	Pass	1	19
	Fail	19	1
	Pc = 38/40 = .95		
4	Pass	1	14
	Fail	19	6
	Pc = 33/40 = .825		
5	Pass	0	4
	Fail	20	16
	Pc = 24/40 = .60		

"true score types" for any given domain: masters, (M); and nonmasters, ($\bar{M}$). Masters are those individuals who have acquired the necessary skills to respond correctly to all items within the domain. Thus for a three-item test with items sampled from the domain of interest, a master's true score response pattern would be 111, where a "one" indicates a correct response to an item. Conversely, nonmasters have not acquired the necessary skills to respond correctly to any item within the domain; thus their true score response pattern would be 000, where a "zero" indicates an incorrect response to an item. This dichotomous classification of individuals appears reasonable to the degree that all items within a domain involve the same skill.

In general, it is assumed that the only way that any non-true score response pattern can occur is for a nonmaster to make one or more correct "guessing" errors or for a master to make one or more forgetting errors. For the first model (Macready & Dayton, 1975), the error probabilities are unrestricted except for the usual 0, 1 bounds for probabilities. a_i and b_i represent the probabilities of a "guessing" and "forgetting" error, respectively, for item i. Furthermore, P(M) and $P(\bar{M})$ represent the proportions of examinees who are masters and nonmasters, respectively, with the usual restrictions: $0 < P(M) < 1$ and $P(M) + P(\bar{M}) = 1$. If local independence among responses is assumed, then the probability of the jth observed response pattern on an n-item test is

$$p(j) = p(j|\bar{M})p(\bar{M}) + p(j|M)p(M)$$

$$= \left[\prod_{i=1}^{n} a_i^{x_{ij}} (1 - a_i)^{1 - x_{ij}} \right] p(\bar{M}) +$$

$$\left[\prod_{i=1}^{n} b_i^{1 - x_{ij}} (1 - b_i)^{x_{ij}} \right] p(M) , \qquad (1)$$

where $x_{ij} = [0,1]$ is the score of the ith item for the jth response pattern. Maximum likelihood estimates of these parameters are obtained from test data by means of the Newton-Raphson iteration procedure (Rao, 1965, pp. 302-309).

Because of the relatively large number of parameters (2n + 1) under this first model, there are circumstances in which it is desirable to utilize a second model (Dayton & Macready, 1976) based on a more restrictive set of assumptions; guessing errors for all items are equal (i.e., $a_i = a$) and "forgetting" errors for all items are equal (i.e., $b_i = b$). These assumptions reduce the number of parameters to be estimated to three for tests composed of any number of items and allow for a

simplification of the formula defining the probability of the occurrence of the jth response pattern on an n-item test to

$$p(j) = p(j|\bar{M}) + p(j|M) = a^{s_j}(1-a)^{n-s_j}p(\bar{M}) + b^{n-s_j}(1-b)^{s_j}p(M), \quad (2)$$

where s_j is the number of correct responses (i.e., number of 1's) in the response pattern.

Macready and Dayton provide a discussion of how these models can be used for making classification decisions with respect to mastery of specific concepts or skills, and they provide several examples. The discussion includes the development of procedures for (1) assessing the adequacy of "fit" provided by the models, (2) identifying optimal decision rules for mastery classification that incorporate utility functions related to costs of false negatives and false positives, and (3) identifying minimally sufficient numbers of items necessary to obtain acceptable levels of misclassification.

Example. For the case of a three-item test, there are eight possible response patterns: (000), (001), (010), (100), (110), (101), (011), (111). For the first model, the 2n + 1 necessary parameters correspond to guessing (a_i) and forgetting (b_i) parameters for each item and the proportion of subjects in the examinee group who are masters. Maximum likelihood estimates of these parameters are obtained from the test data.

For purposes of example for Model I, assume the following parameter values: $a_1 = .01$, $b_1 = .20$; $a_2 = .05$, $b_2 = .10$; $a_3 = .10$, $b_3 = .05$; and $P(M) = P(\bar{M}) = .5$. This might correspond to a test in which the items appeared to be growing increasingly easy. For the second model, only three parameters are found: a, b, and P(M). Again for purposes of example for Model II, assume that the obtained estimates for the parameters are $a = .06$, $b = .12$, $P(M) = P(\bar{M}) = .5$.

To find the probability of observing each response pattern in a given examinee group, the probability of observing each response pattern given mastery status must be multiplied by the proportion of the group in that mastery status. For this example, each response pattern must be multiplied by $p(M) = P(\bar{M}) = .5$. Table 3 shows the results of these calculations.

The mastery/nonmastery decision rule is based on the score that minimizes the probability of misclassification. Probability of misclassification is defined as the probability that a master will not achieve the cutting score times the proportion of masters in the group

Table 3

Probability of Observing Response Patterns Under the Macready and Dayton Models, Assuming $P(M) = P(\bar{M}) = .5$

Response pattern	Model I		Model II	
	P(response pattern)		P(response pattern)	
	Master	Nonmaster	Master	Nonmaster
000	.0005	.423225	.000864	.415292
001	.0095	.047025	.006336	.026508
010	.00450	.022275	.006336[c]	.026508[d]
100	.0020[a]	.004725[b]		
110	.0180	.000225	.046464	.001692
101	.0380	.000475	.046464	.001692
011	.0855	.002475	.046464	.001692
111	.3420	.000025	.340736	.000108
	$P(M) = .5$	$P(\bar{M}) = .5$	$P(M) = .5$	$P(\bar{M}) = .5$

[a] $p(M) = (.2^0 \times .8^1)\ (.1^0 \times .9^1)\ (.05^1 \times .95^0) \times .5 = .0180.$

[b] $p(\bar{M}) = (.01^1 \times .99^0)\ (.05^1 \times .95^0)\ (.1^0 \times .9^1) \times .5 = .000225.$

[c] $p(M) = .12^2 \times .88^1 \times .5 = .006336.$

[d] $p(\bar{M}) = .06^1 \times .94^2 \times .5 = .026508.$

plus the probability that a nonmaster will equal or exceed it times the proportion of nonmasters in the group. The probabilities for both models and all possible cutting scores are given in Table 4.

The final column of Table 4 indicates that for both models the optimal cutting score is 2 correct. Note that although the cutting score is the same for both models, the misclassification under the richer Model I is consistently smaller than Model II.

Emrick (1971) developed a procedure related to the restricted form of the Macready and Dayton model. He generated a function for identifying optimal cutoff scores in terms of relative costs of incorrect mastery/nonmastery decisions and the ratio of a to b errors. The optimized formula is

$$k = \frac{\log \frac{b}{1 - a} + \frac{1}{n} \log \left[\frac{L_2 P(M)}{L_1 P(\bar{M})}\right]}{\log \frac{ab}{(1 - a)(1 - b)}}, \qquad (3)$$

where

k = percentage of items correct required for a mastery decision;
L_1 = loss incurred from a false positive;
L_2 = loss incurred from a false negative.

This cutscore value is the same as that suggested by Macready and Dayton under their restricted model when the same parameter estimates are used. However, Emrick suggests a different approach for parameter estimation. He constructs a fourfold table relating true mastery state and observed item responses to a single item, with the cell entries being the error probabilities a and b. Emrick then treats a and b as response contingencies and computes a phi coefficient to indicate the correlation between observed single item responses and true mastery state:

$$\text{phi} = \frac{1 - a - b}{\sqrt{1 - (a - b)^2}}. \qquad (4)$$

He uses the average iteritem correlation of examinee responses to compute an unbiased estimate of the reliability of a single item using the Spearman-Brown prophecy formula.

Since reliability is defined as the proportion of total variance that is true variance, it can be interpreted as an unbiased estimate of the squared correlation between an examinee's true mastery state and his

Table 4

Probability of Misclassification as a Function of Cutting Score Under the Macready and Dayton Models, Assuming $P(M) = P(\bar{M}) = .5$

Cutting score	P(False negative)		P(False positive)		P(Misclassification)	
	Model I	Model II	Model I	Model II	Model I	Model II
0 (all pass)	0	0	.5	.5	.5	.5
1	.0005	.000864	.076775	.084708	.077275	.085572
2	.01650[a]	.019872[b]	.0032[c]	.005184[d]	.0197	.025056
3	.1580	.159264	.000025	.000108	.158025	.159372
4 (all fail)	.5	.5	0	0	.5	.5

[a], [b] The probability that a master will be misclassified when the cutoff score is set at 2 correct equals the sum of the probabilities that a master will get only 0 or 1 items correct times the proportion of masters in the group. For Model I, this probability equals .0005 + .0095 + .0045 + .002 = .0165. For Model II, .000864 + 3(.006336) = .019872.

[c], [d] The probability that a nonmaster will be misclassified when the cutoff score is set at 2 correct equals the sum of the probabilities that a nonmaster will get 2 or 3 items correct times the proportion of nonmasters in the group. For Model I, this probability equals .000025 + .002475 + .000475 + .000225 = .0032. For Model II, .000108 + 3(.001692) = .005184.

or her item response. Hence, item responses, true mastery state, and error probabilities can be directly related through the test reliability. If the ratio of a to b is known (or if it can be estimated), values for a and b can be directly calculated.

For the Macready-Dayton model example values (a = .06, b = .12), the value of phi is .821. Squaring this value and applying the Spearman-Brown prophecy formula for a three-item test indicates that the test reliability for this example would be .86. Assuming a loss ratio of 1 and equal proportions of masters and nonmasters, the value for k in Emrick's optimization formula is .4339. This implies a cutting score of 1.3 on a three-item test, or rounding up to the next higher integer, 2. Thus, the final result is the same as the result obtained with Macready and Dayton.

Evaluation. An important constraint of this approach is that the proportion of masters and nonmasters must be equal. (The computations for the preceding example and a more general form of the Emrick model are presented in Appendix A.)

Other possible weaknesses in Emrick's approach to parameter estimation are the subjectivity required and the somewhat overly restrictive assumptions necessary to implement his approach. In addition, the complexity of both conceptualizing and quantifying L_1 and L_2 may greatly complicate the derivation of cutoff scores under these models.

If the assumptions are met, an optimal differentiation between masters and nonmasters will result. Furthermore, a means is provided to determine how many items are needed to keep the probability of misclassification at or below some specified critical level. The relationships among test items may also be explored. A major potential weakness concerns the assumption that learning occurs in an "all-or-none" manner, with no partial learning or overlearning. Failure to satisfy this assumption could produce a poor fit of data to the model, which will in turn produce a far less than optimal cutting score.

Binomial Model

Assumptions and Rationale. In contrast to the all-or-none learning assumption of the Emrick and Macready models is the assumption that learning is a continuous process. A binomial distribution model, first suggested and derived by Kriewall (1969) and subsequently developed by Millman (1972), defines proficiency as the probability that a person will correctly respond to any test item randomly chosen from a specified domain of items. Proficiency may also be defined as the proportion of items that would be correct if all items in the domain could be administered. Since the proficiency value can take on values from zero to one, the model allows for partial acquisition.

The following assumptions are pertinent: (1) dichotomously scorable items, (2) local independence of items, (3) no systematic learning or forgetting during test taking, and (4) items equally difficult for any given examinee. The percentage of items answered correctly is taken as a point estimate of the examinee's true proficiency. For a given proficiency, the probability of observing any score may be determined. The hypothesis to be tested in this model involves the likelihood of a specific score, if indeed the examinee had the given level of proficiency.

The basic equation for the binomial model yields the probability distribution of scores for an examinee with proficiency "p" for repeated random samples of items of size "n" from a given domain of items:

$$f(x) = \binom{n}{x} p^{x} (1 - p)^{n - x} , \qquad (5)$$

where

x = the total number of correct responses,
$f(x)$ = the probability of test score x,
$\binom{n}{x}$ = the binomial coefficient:

$$\frac{n!}{x!\,(n - x)!} .$$

The binomial model can be used to provide two types of information. First, the proportion correct is the maximum likelihood estimate of an individual's proficiency relative to the particular domain. Second, the model can be used to investigate the interaction between test length and classification error when individuals are divided into two groups. One group will contain students with proficiency greater than or equal to some minimal proficiency criterion. The other group will have students with proficiency levels less than or equal to some maximum nonmastery criterion.

To calculate the expected error in decisionmaking, it is necessary to specify two parameters. The first is the lowest proficiency level required for an individual to be considered a master. The second is the highest proficiency level that a student could obtain and still be considered a nonmaster. When these values are set by the decisionmaker, the probability of false negative and false positive errors for minimal masters and maximal nonmasters, respectively, can be calculated for any given test length and cutting score. This procedure, it should be noted, is generally conservative. That is, if the group contains examinees with abilities above minimal mastery or below maximal nonmastery, the number of misclassifications observed will be less than that predicted by the model.

Example. Suppose that a cutoff score of 80% correct was selected (i.e., in order to be classified as a master, a student must get correct at least 80% of whatever number of items are included on the test). Assume also that a true proficiency of 90% is defined as the minimal mastery level, and that a true proficiency of 70% is defined as the maximal nonmastery level. The region between these cutoff scores is an "area of indifference." That is, if an examinee's true proficiency lies between 70% and 90%, the decisionmaker would be indifferent as to whether the examinee is classified as a master or as a nonmaster.

Values for misclassification error that can be tolerated must also be specified. Continuing with the above example, assume that the decisionmaker is unwilling to accept more than 26% of the students whose true ability is 70%, and he or she wants to reject not more than 19% of those whose true ability is 90%. Thus, the probabilities of a false positive and false negative are .26 and .19, respectively. Given these values, it is possible to determine the minimal number of test items.

The following notation will be used:

n = the total number of test items,
c = the cutoff score (in this example c = .8n or the next highest integer value of .8n since an 80% standard was chosen),
x = the observed score, and the formula for cumulative terms of the binomial distribution is

$$\sum_{x=1}^{n} \binom{n}{x} p^x (1-p)^{n-x} . \qquad (6)$$

Specifying that the probability of falsely rejecting a master must not exceed .19 means that the cumulative probability of a master obtaining a score from 0 correct to c - 1 correct must not exceed .19. This constraint may be expressed as the inequality

$$F(x \leq c - 1) \leq .19. \qquad (7)$$

Therefore,

$$.19 \leq \sum_{c=0}^{x=c-1} \binom{n}{x} (.9)^x (.1)^{n-x} ,$$

where p = .9, the minimal mastery level.

A similar relationship exists for nonmasters. Since the probability of falsely accepting a nonmaster must not exceed .26, the cumulative

probability of a nonmaster obtaining a score greater than or equal to c must not need exceed .26. The inequality for nonmasters is

$$F(x \geq c) \leq .26. \tag{8}$$

Therefore,

$$.26 \leq \sum_{x=c}^{n} \binom{n}{x} (.7)^{x} (.3)^{n-x},$$

where p = .7, the maximal nonmastery level.

Reference to a table of cumulative terms of the binomial distribution shows that the minimum value of n for which these relationships hold is 8.

Since .8 (8) = 6.4, a cutoff score of 7 correct is chosen. Substituting these values for c and n yields

$$.19 = \sum_{x=0}^{x=6} \binom{8}{x} (.9)^{x} (.1)^{8-x} \quad \text{and} \tag{9}$$

$$.26 = \sum_{x=7}^{x=8} \binom{8}{x} (.7)^{x} (.3)^{8-x}. \tag{10}$$

These are the numerical solutions for the above inequalities.

The conservative nature of the model results from the fact that the calculations are based on two point values of true proficiency, 70% and 90%. The previous calculations reflect the probabilities of false-positives and false negatives, assuming that the examinee group is composed only of people with true proficiencies of 70% and 90%. However, if an examinee had a true proficiency of 95%, the probability that he or she would obtain a score of less than seven correct out of eight items, and therefore be classified as a nonmaster, may be expressed as

$$\sum_{x=0}^{x=6} \binom{8}{x} (.95)^{x} (.05)^{8-x} = .06. \tag{11}$$

This value is considerably less than the probability of a false negative as previously obtained, .19.

On the other hand, if a person had a true proficiency equal to 60%, the probability that he or she would obtain a score of seven or more

correct on an eight-item test, and therefore be classified as a master, may be expressed as

$$\sum_{x=7}^{x=8} \binom{8}{x} (.6)^{x} (.4)^{8-x} = .11. \qquad (12)$$

This value is much less than the probability of a false positive as previously obtained, .26.

Millman (1972) has prepared tables which allow the decisionmaker to reach these same conclusions without calculations. His tables also give the expected misclassification error for a variety of test lengths, cutoff percentages, and true ability levels.

Evaluation. The binomial model actually describes the worst possible situation. For most practical applications, the examinee population will contain persons with true ability above the minimal mastery level and below the maximal nonmastery level. To arrive at a more realistic estimate of total misclassification, the equations would have to be solved for each representative ability and be weighted by the proportion of the group with each ability. Such a procedure is, of course, feasible but its value is questionable. The values obtained from the simple procedure are overly pessimistic; any decision derived from empirical data could be no worse, and would probably be better.

A virtue of this model is that it is relatively straightforward, being based on the familiar binomial distribution. It is one of the simpler quantitative models to derive test lengths and cutting scores. The model can be criticized, however, because of its conceptual foundations. Specifically, the output of the model tells us the probability that a student will attain a certain test score, given his or her true ability level. However, it is by no means clear or obvious that the decisionmaker would know the student's true level of functioning. Indeed, if the true ability level were known, there would be no need for models to determine test length and cutting scores. In using the binomial model, the decisionmaker has to set estimated (or desired) limits on the true level of functioning of the student. This allows him or her to infer the conditional probability of the observed test score, given the hypothesized level(s) of proficiency. This binomial model is most useful for initial approximations of test length and cutting score before test data have been collected.

Bayesian Model

Assumptions and Rationale. If information can be obtained about the quality of the examinee population (perhaps on the basis of previous similar populations) before the test scores are observed, then a Bayesian model may be appropriate for deriving test lengths and cutting

scores. The input consists of an estimate of the ability distribution in the examinee population, and the conditional probabilities that a randomly chosen item would be answered correctly given some ability level. The output is the conditional probability that an individual's ability equals (or, in some cases, exceeds) some criterion ability, conditional upon his or her test score.

The Bayesian, like the binomial, model makes the following assumptions: (1) items must be dichotomously scored, (2) responses are independent, (3) items are equally difficult for any given examinee within a particular ability group, and (4) there is no systematic learning or fatigue during test taking. As in the binomial model, ability is defined as the probability of responding correctly to a randomly chosen item from the domain. We will continue to use the term proficiency (p) when referring to this definition of ability.

Examples. The first model to be discussed assumes $i \geq 2$ discrete states of mastery.

Epstein and Steinheiser (1975) developed a two-step algorithm based on work by Hershman (1971). The first step yields the probability of an examinee being in mastery state i, conditional on an item score:

$$p(M_i|t) = \frac{p(t|M_i)\ p(M_i)}{\sum_{i=1}^{s} p(t|M_i)\ p(M_i)}, \qquad (13)$$

where s = the number of states,
t = the item score (0 or 1),
M_i = the mastery state being considered,
$p(M_i)$ = the prior probability that an individual is in mastery state i, and
$p(t|M_i)$ = the probability of the score t, given the mastery state.

The second step in the procedure combines the decisions for each item into a final probability of being in mastery state i, given the total test score:

$$p(M_i|T) = \frac{\prod_{j=1}^{n} p(M_i|t_j)}{p(M_i)^{n-1} \sum_{i=1}^{s} \left[\frac{\prod_{j=1}^{n} p(M_i|t_j)}{p(M_i)^{n-1}} \right]}, \qquad (14)$$

where

$j = 1, 2, \ldots n$ = the number of items and
T = the total test score.

For example, consider the case previously described for the binomial model. Two mastery states are assumed, minimal mastery and maximal nonmastery.

For the minimal mastery state (M_1), $p(t_j = \text{correct } (1) \mid M_1) = .9$ and $p(t_j = \text{incorrect } (0) \mid M_1) = .1$, for all j.

For the maximal nonmastery state (M_2), $p(t_j = \text{correct } (1) \mid M_2) = .7$ and $p(t_j = \text{incorrect } (0) \mid M_2) = .3$.

Values must be given for the priors, $p(M_1)$ and $p(M_2)$. Their value may be determined on the basis of past experience, or may simply reflect the beliefs or expectations of the evaluator. Three cases will be considered: $p(M_1) = p(M_2) = .5$; $p(M_1) = .12$, $p(M_2) = .88$; and $p(M_1) = .62$, $p(M_2) = .38$. These correspond to little prior information, relatively low expectations, and relatively high expectations. The example was computed for an observed score of seven correct on an eight-item test. The results are shown in Table 5.

For Cases 2 and 3, where prior information favored the nonmastery and mastery states, the final decision can be made with a relatively high degree of confidence. For the case of little prior information, Case 1, the probabilities of misclassification are greater. The effects on the final decision of the priors are also clear. For the equal priors case, the weight of the observed evidence favors a mastery decision. However, where the nonmastery state is favored in the prior probabilities (Case 2), the evidence does not overcome the priors and a nonmastery decision is made.

Whereas the Epstein and Steinheiser technique seems to offer a method for reducing the uncertainty in decisionmaking for a given number of test items, their procedure is limited by the constraint that only discrete mastery groups are considered. The second model to be reviewed deals with continuous distributions of proficiency and classifies examinees based upon the probability that their proficiency equals or exceeds some minimal criterion. Novick and Lewis (1974) achieve this by assuming that the distribution of examinee proficiencies can be approximated by a member of the family of Beta distributions. The probability of achieving any score of interest, given the proficiency, remains binomial. The form of Bayes' Theorem is then a probability density function of the form $p(T|x) = p(x|T)p(T)$, where T is the proficiency and x is the test score.

If $p(x|T)$ is binomial and $p(T)$ is a Beta distribution, then $p(T|x)$ will also be a member of the Beta family. In fact, if the prior

Table 5

Changes in Posterior Probability of Mastery as a Function of Changes in Prior Probability of Mastery

	Prior		
$p(M_1)$	.5	.12*	.62
$p(M_2)$	.5	.88	.38
	Posterior		
$p(M_1 \mid T)$	.66	.205*	.767
$p(M_2 \mid T)$	.33	.796	.242

*Computational steps: $p(t_j = 1) = .12 \times .9 + .88 \times .7 = .724$

$p(t_j = 0) = .12 \times .1 + .88 \times .3 = .276$

$p(M_1 \mid t_j = 1) = (.12 \times .9)/.724 = .149$

$p(M_1 \mid t_j = 0) = (.12 \times .1)/.276 = .043$

$\Pi p(M_1 \mid t_j) = .149^7(.043) = 7 \times 10^{-8}$

$p(M_1 \mid T) = 7 \times 10^{-8}/[(.12^7)(7 \times 10^{-8}/.12^7 + .309/.88^7)] = .205$

$p(M_2 \mid T) = .309/[.88^7(7/36 + .755)] = .796$

$p(M_2 \mid t_j = 1) = (.88 \times .7)/.724 = .851$

$p(M_2 \mid t_j = 0) = (.88 \times .3)/.276 = .957$

$\Pi p(M_2 \mid t_j) = .851^7(.956) = .309$

distribution is Beta (a,b) (i.e., B(a,b)), and a score of x is observed in n trials, then the posterior distribution is B(a + x, b + n - x).

Continuing with the previous example in the continuous framework, we shall now consider three prior distributions. Integer values of a and b, the parameters of the Beta distribution, will be used. We may therefore use the Incomplete Beta function $I_p(a, b)$, which has the following relationship to the cumulative binomial distribution:

$$\sum_{x = x'}^{n} \binom{n}{x} p^x q^{n - x} = I_p(x', n - x' + 1), \qquad (15)$$

where n is the number of test trials, p is the probability of success on a randomly selected trial, and x' is the observed number of successes.

Tabled values are available (Beyer, 1966, Table III.2). For non-integer values of a and b, programed numerical methods may be required (Novick & Jackson, 1974).

For the first example, assume that little is known about the examinee population, i.e., a randomly selected examinee may get a test score that would place him or her in the mastery or nonmastery category with equal probability. In terms of the Beta distribution, this means that examinee proficiency would be rectangularly distributed, resulting in a = 1, b = 1, or B(1, 1) (Novick & Jackson, 1974, p. 114).

For the second case, assume that the prior probability that a randomly chosen examinee has proficiency greater than or equal to .8 is .12, i.e., $P(\tilde{p} \geq .8) = .12$. Therefore, $1 - \tilde{p}$ is used to enter the cumulative binomial table at the top (since tabled $\tilde{p}$ values stop at $\tilde{p} = .50$), and .12 is the table value.

However, we cannot use the table until one more parameter is specified; so let us assume that the examiner's "certainty of prior belief" can be quantified as being equivalent to the information that would be available if a 10-item test were given (Winkler, 1972, p. 187). With n = 10, we find that an entry with a value of .12 in the .20 column for n = 10 has an associated x' value equal to 4. Unfortunately, x' does not equal 4, due mainly to a limitation of the table, since $\tilde{p}$ values stop at .50 and do not extend to .80 or beyond. Note, however, that if we let x' = 4 in the cumulative binomial, and subtract the result from 1, we obtain

$$\sum_{4}^{10} \binom{n}{x} (.2)^x (.8)^{n - x}, \text{ which equals } 1 - .1208, \text{ or } .88.$$

If the table had extended to $\tilde{p} = .8$, then the value .879 would have been found as the entry corresponding to n = 10 and x' = 7. Hence, the value for x' is 7. Substituting x' = 7 and n = 10 in equation (15), we obtain $I_p(7, 4)$ as the Beta distribution which represents the prior information that $P(\tilde{p} \geq .8) = .12$ is equivalent to 10 additional test trials.

The third example considers that the prior probability of a randomly chosen examinee having proficiency greater than or equal to .8 is .62--which is also comparable to information that could be obtained from a 10-item test. Again, entering the table with n = 10, $1 - \tilde{p} = .2$, we find that a tabled value of .62 this time corresponds to x' = 2. Substituting x' = 2 in the cumulative binomial and subtracting that result from 1 yields .38. Again, an extension of the table to $\tilde{p} = .8$ would show that when n = 10, a tabled value of .38 corresponds to an x' value of 9. Therefore, the parameters for the Beta distribution in this case are $I_p(9, 2)$.

Having thus derived the prior distributions, let us now consider some hypothetical test scores, and then derive the posterior distributions.

Suppose that a score of seven correct on an eight-item test were observed. Then the posterior proficiency distributions will be B(a + number correct, b + number of trials - number correct). For the three examples, we therefore have B(8, 2), B(14, 5), and B(16, 3).

The posterior probability that an examinee with a score of seven correct out of eight items has a proficiency greater than or equal to .8 (i.e., $P(\tilde{p} \geq .8 \mid 7, 8)$) can be found by determining the area in the upper tail of the appropriate Incomplete Beta function (Winkler, 1972, Table 5; Schlaifer, 1969, Table T3; Novick & Jackson, 1974, Table A-14). For the three examples, these values are: $I_{.8}(8, 2) = .56$; $I_{.8}(14, 5) = .28$; and $I_{.8}(16, 3) = .73$.

Since the origin of these values may not be intuitively obvious, we shall outline the steps required to complete the first example, using the Novick and Jackson tables.

Step 1: Since p > q, reverse the order, and enter the table with p = 2 and q = 8.

Step 2: The table gives the cumulative area (of proficiency); however, since we want to determine the area in the upper part of the Beta function, we need to subtract the stated proficiency of .8 from 1, and thereby obtain .2. This represents the symmetric area in the lower 20% of the distribution.

Step 3: .2 lies between the tabled values of .1796 and .2723, with associated probabilities (fractiles) of those tabled proficiencies equal to 50% and 75%, respectively.

Step 4: Interpolation yields the fact that a 20% or less proficiency would occur 56% of the time; therefore, 80% or greater proficiency should also be observed 56% of the time.

Novick and Jackson also provide a convenient set of charts (pp. 122-123) for rapid approximations, although it should be noted that for the current example, the solution is found to be .44 from their chart A. This value must be subtracted from 1, since the .44 represents the cumulative area in the lower portion of the B(8, 2) curve.

If the probability of having a proficiency greater than or equal to .8 must be at least .5 for an examinee to be classified as a master, then a score of 7 out of 8 would lead to a mastery classification only in the first and third examples previously described. The weight of the low prior reversed the decision rule in the second example.

For another approach to deriving prior distributions, assume that prior information can be described as equivalent to 7 correct on a 10-item test. (This is an assumption not without criticism, as we shall note in a subsequent section.) Assume also that proficiency is distributed as Beta--a helpful and reasonably appropriate assumption. The mean of the examinees' proficiency then equals $(x/n + 1)$ or $7/11 = .636$. The variance equals $x(n - x + 1)/(n + 1)^2(n + 2) = 28/1452 = .019$. Since the parameters are integers, we may once again use the cumulative binomial as a means of obtaining the Incomplete Beta density function:

$$I_p(7, 4) = \sum_{x=7}^{n=10} \binom{10}{x} p^x q^{10-x} \qquad (16)$$

$$\frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^x u^{a-1}(1-u)^{b-1}du.$$

Equation (16) is the probability that a given proficiency is less than or equal to p. We can compute this probability by assigning specific values to p, as shown in Table 6. The values for $P(p \geq p)$ up to the 50th fractile may be found directly (Beyer, 1966, Table III.2) for $x' = 7$ and $n = 10$. Values for .6 and greater can be computed according to the cumulative binomial equation (16). When the values obtained (as in Table 6) are plotted, the result is a smooth ogive-like curve (Winkler, 1972, pp. 153, 186; Schlaifer, 1969, p. 438).

To plot the proficiency distribution, we may use the Beta distribution function:

$$\frac{\Gamma(a+b)u^{a-1}(1-u)^{b-1}}{\Gamma(a)\Gamma(b)} \qquad (17)$$

Table 6

Cumulative Estimation of Prior Probabilities for Various Assumed Proficiencies

p = Proficiency	$I_p(7,4)$, or $P(\tilde{p} \leq p)$
.1	.0000
.2	.0009
.3	.0106
.4	.0548
.5	.1719
.6	.3823
.7	.6496
.8	.8791
.9	.9872

Values of the proficiency (p) may be chosen, but $a = x' = 7$, and $b = n - x' + 1 = 4$. Since $(n + 1) = n!$ for integers, we can easily solve equation (16): $\Gamma(a + b) = \Gamma(11) = 10! = 3.6288 \times 10^6$; $\Gamma(a) = \Gamma(7) = 6! = 7.2 \times 10^2$; $\Gamma(b) = \Gamma(4) = 3! = 6$. Therefore, $\Gamma(a + b)/\Gamma(a)\,\Gamma(b) = 3.6288 \times 10^6/(720)(6) = 840$. Table 7 shows how values of f(p) may be obtained.

A plot of the tabled values for p on the abscissa and f(p) on the ordinate could then be made. Such plots may also be found in Winkler (1972, sec. 4.3 and 4.4), Schlaifer (1969, sec. 11.1.2) and Novick and Jackson (1974, p. 112). Note that this is a prior distribution of hypothesized proficiencies in which we assumed at the outset that the information could be characterized as comparable to the information that would be obtained from observing a score of seven correct on a ten-item test.

Evaluation. Bayesian models offer the possibility of enhancing the assessment of examinee proficiency by using prior information, e.g., knowledge that content experts or examiners have about previous similar examinee populations. As the validity and accuracy of this prior information increases, fewer test items will be needed to achieve a given level of classification accuracy in comparison to the binomial model and in comparison to the Bayesian case of equal priors. As more is known about the examinee population (i.e., the more that prior information departs from a B(1, 1) distribution), the more the variability in the posterior distribution is reduced, and the more the number of items to attain a desired level of accuracy is reduced.

In comparing the binomial and Bayesian models, note that the former produced as output the probability of observing a specific score conditional upon some hypothesized true ability level. In the spirit of classical hypothesis testing, one need not know anything about an examinee's proficiency, except that he or she is more or less likely to come from the mastery side of the cutoff score. Since some true level of functioning must be hypothesized, it is possible to determine the probabilities of falsely passing a nonmaster and falsely failing a master if the test score suggests a true proficiency level either above or below the hypothesized true level of functioning.

In contrast, the Bayesian model provides as output the probability that a specific examinee has a true ability equal to or greater than the criterion (minimal) ability, conditional upon the observed test score. But since no true ability was hypothesized, false positive and false negative error rates cannot be specified as was possible with the binomial model. While both models give the probability that an examinee is a member of some ability level group, the binomial estimate refers to the probability of a score occurring conditional upon the assumed true proficiency; whereas the Bayesian estimate refers to the probability of a specific examinee being at or beyond some proficiency level conditional upon his or her observed test score.

Table 7

Point Values for Prior Proficiency Distribution

Proficiency values	$p^{a-1}(1-p)^{b-1}$	$f(p) = 840(p)^{a-1}(1-p)^{b-1}$
.1	7.29×10^{-7}	6.12×10^{-4}
.2	3.28×10^{-5}	2.75×10^{-2}
.3	2.50×10^{-4}	2.10×10^{-1}
.4	8.85×10^{-4}	7.44×10^{-1}
.5	1.95×10^{-3}	1.64×10^{0}
.6	2.99×10^{-3}	2.51×10^{0}
.7	3.18×10^{-3}	2.67×10^{0}
.8	2.10×10^{-3}	1.76×10^{0}
.9	5.31×10^{-4}	4.48×10^{-1}

There are several difficulties confronting the potential user of a Bayesian model for CRT purposes. First, the mathematics can become rather cumbersome, since the Beta distribution must be used when ability is assumed to be distributed continuously. Second, a methodological difficulty arises in the determination of prior probabilities (Winkler, 1972, sec. 4.8). It is methodologically unsound to merely ask the examiner or expert to "state his priors," since simple human judgment of probabilities is often unreliable, inconsistent, and distorted (Kaplan & Schwartz, 1975). A method used in the present paper--equating prior information to comparable test length and score information--may be suitable for purposes of illustration, but it may be difficult to implement in applied settings.

There is at present a dearth of research about how prior probabilities can actually be obtained from experts. Perhaps a pair comparison or forced-choice procedure could be used in which various combinations of proficiency (or expected scores) and associated probabilities are presented to the expert (Steinheiser, 1976). Thus, the judge's prior distribution would be directly obtained, and the best fitting Beta distribution used to provide the necessary parameter values.

Rasch's One-Parameter Logistic Model

Assumptions and Rationale. The latent trait model developed by Rasch (1960, 1961, 1966) is claimed to yield person-free test calibrations and item-free person measurements (Wright & Panchapakesan, 1969). The model attempts to reproduce an item by score group matrix in which n items are ordered by their difficulties, and n - 1 score groups are ordered by the raw scores. Cell entries represent the probability that item i will be passed by a person in score group j (Whitely & Dawis, 1974).

There are two parameters in the model. The first is person ability A; the second is item difficulty D. The odds (O) of a person correctly answering an item are equal to the product of the person's ability times the item's difficulty: $O = A \times D$. If we express the odds as a probability, we find that the probability P of a person with ability A succeeding on an item with difficulty D can be expressed as

$$P = \frac{A \times D}{1 + A \times D} .$$

Replacing A and D with their logarithms, log A = a and log D = d, we may finally express P as a logistic function (Wright, 1967):

$$P = \frac{1}{1 + e^{(-a - d)}} . \qquad (18)$$

This model assumes that (1) all items measure the same unidimensional trait; (2) all items have equal discriminating power and vary

only in difficulty (the restriction of a common discrimination index results in a set of nonintersecting item characteristic curves which differ only by a translation along the ability scale); (3) subjects and items are locally independent; (4) guessing effects are negligible, and (5) there is no time constraint on answering items (Rasch, 1966).

Tests comprised of items all of which fit the model have the following properties (Wright & Panchapakesan, 1969; Whitely & Dawis, 1974): (1) estimates of item difficulty parameters will not differ significantly for any sample of examinees; (2) estimates of person ability will not differ significantly for any sample of calibrated items; (3) individual ability estimates can be measured on at least an interval, and perhaps a ratio scale (Wright, 1967); (4) the scale of abilities is defined regardless of the characteristics of the subject population who take the test; and (5) a unique standard error of measurement is associated with each ability level.

The significance of the Rasch logistic model may be appreciated by comparing it to "classical" models of test development:

> A psychological test having these general characteristics would become directly analogous to a yardstick that measures the length of objects. That is, the intervals on the yardstick are independent of the length of the objects, and the length of individual objects is interpretable without respect to which particular yardstick is used. In contrast, tests developed according to the classical model have neither characteristic. The score obtained by a person is not interpretable without referring to both some norm group and the particular test form used. . . . No longer would equivalent forms need to be carefully developed, since measurement is instrument independent and any two subsets of the calibrated item pool could be used as alternative instruments. Similarly, independence of measurement from a particular population distribution implies that tests can be used for persons dissimilar from the standardization population without the necessity of collecting new norms (Whitely & Dawis, 1974, 163-164).

Examples. Calibrating a test using the Rasch model results in a logarithmic ability estimate being assigned to every possible raw score. This estimate indicates the amount of ability required to achieve that raw score. A comparison of the ability estimates assigned to a given raw score by two samples with different ability distributions indicates the degree to which the Rasch model calibrates a test independently of the ability level of the calibration sample.

Wright (1967) studied the responses of 976 beginning law students to 48 reading comprehension items on the L.S.A.T. To obtain samples with different ability distributions, he selected two contrasting

groups from his total sample. The lower group included the 325 students who did poorest on the test, with a top score of 23. The higher group included the 303 students with the highest scores, with a bottom score of 33. Wright compared the similarity between the two sets of Rasch ability estimates and the two sets of percentile ranks. Figure 1 shows the results, in terms of "person-bound test calibration," where a plot of raw score against percentile rank clearly shows two different ability groups. If a person is said to be in the nth percentile, reference must be made to which group that person belongs.

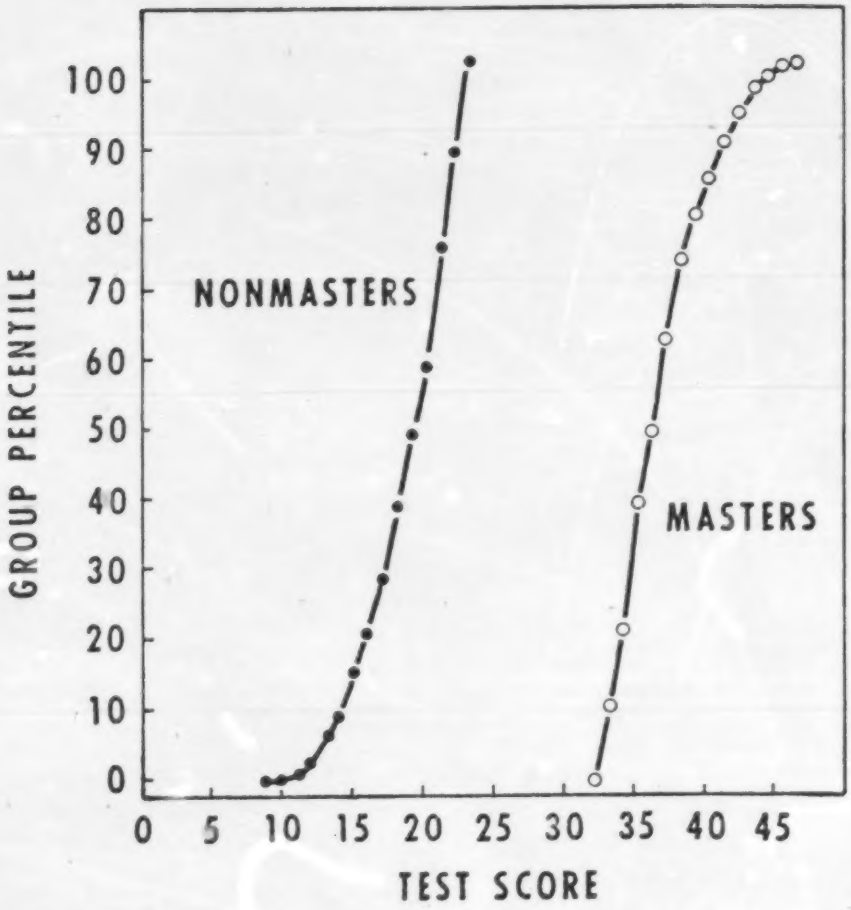
After subjecting these same data to the Rasch logistic analysis, the test scores are transformed into ability measurements along the ordinate. Figure 2 shows that the curves for the best and worst examinees almost completely overlap.

The difficulty estimates based upon these dichotomous examinee groups are statistically equivalent. Therefore, these estimates are independent of the ability of the examinees in the calibration sample, and may be used over the entire range of ability. Comparing the calibration curves of these figures shows the contrast between (1) calibration based upon the ability distribution of a standardizing sample, and (2) calibration that is free from the effects of the ability distribution of the examinees used for the calibration.

Can ability be measured in a fashion that frees it from dependence on the use of a fixed set of items? If a pool of test items has been calibrated on a common scale, can any set of items be selected from that pool to make statistically equivalent ability measurements?

Wright (1967) tested these hypotheses by making it as difficult as possible for person measurement to be item free. He divided the original test items into two non-overlapping subtests, the easiest items comprising one subtest and the hardest items comprising the other subtest. The model predicts that ability estimates based upon the easy subtest should be statistically equivalent to those estimates based upon the hard subtest.

The solution required converting the scores to log abilities, and then standardizing the differences in ability estimates. First, for each score, the corresponding log ability on the calibration curves was obtained (see Figure 2). For each pair of scores (from the easy and hard subtests), a pair of estimated log abilities was obtained. Then, a standardized difference was found by dividing the difference between the easy and hard subtest ability estimates by the measurement error of the differences. If the ability estimates are statistically equivalent, then the distribution of standardized differences should have a mean equal to zero and a standard deviation equal to one. The obtained values were .003 and 1.014, respectively.



gure 1. "Person-bound" test calibration using non-overlapping groups of masters and nonmasters (adapted from Wright, 1967).

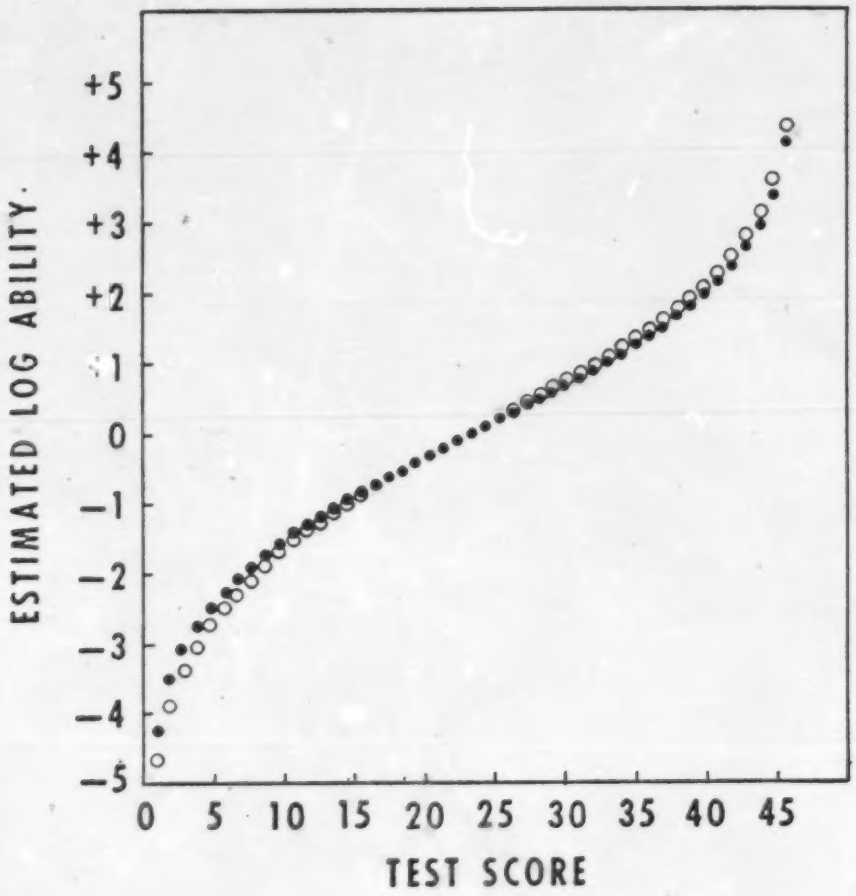


Figure 2. "Person-free" test calibration using the Rasch logistic model ("o" = master, "•" = nonmaster); (adapted from Wright, 1967).

Applications. A more detailed example will show how the Rasch model was used to analyze the results of a criterion-referenced test (Kifer & Bramble, 1974). The data were obtained from 201 college students taking an 84-item multiple choice examination in introductory educational psychology. After discarding items that did not fit the model, the final test contained 68 items.

Comparison of the Rasch-derived ability estimates to a criterion score can proceed in two ways.

The first is analogous to determining the probability of committing a Type I error in classical hypothesis testing. That is, if the criterion ability corresponds to the null hypothesis, we must determine the probability that an obtained ability could have arisen from random sampling from a distribution with a mean equal to the criterion ability and a standard deviation equal to the error associated with the criterion ability.

The second is analogous to determining the probability of committing a Type II error in classical hypothesis testing. That is, given an obtained ability estimate and associated error (standard deviation), we seek the probability that the criterion ability could have been observed from random sampling from the distribution corresponding to the obtained ability estimate.

Kifer and Bramble chose to define their criterion score as 80% of the items correct or 54.4 items correct. Their cutoff score was therefore 55. A raw score of 55 yields an ability estimate of 1.69, with a standard error of .33. Suppose a raw score of 60 were obtained. What is the probability that this score exceeds the criterion score of 55?

The solution requires that we find the probability that this score is part of the criterion distribution, with mean equal to 1.69 and standard devision equal to .33. (1) Kifer and Bramble's parameter estimates show that an observed score of 60 has an ability value equal to 2.32. (2) 2.32 - 1.69 = .63 units of difference between the observed and criterion abilities. (3) .63/.33 = 1.91 standard deviations of difference between the ability values. (4) A table of the normal distribution shows that 1 - F(1.91) = .03. Therefore, the ability value of 2.32 has a probability = .03 of coming from a normal distribution with a mean = 1.69 and standard deviation = .33.

Page 34, para 6, line 3 -- (sp) "deviation" -- not "devision"

There is a second method by which ability estimates may be compared to mastery standards. This method requires the probability that the criterion ability is part of the distribution which has a given (observed) ability as its mean and the given ability standard error as its standard deviation. We now need to find the probability that the true ability corresponding to a score of 60 does not exceed the criterion ability. (1) Kifer and Bramble's parameter estimates show that an observed score of 60 has an ability value equal to 2.32 and a

standard error equal to .39. (2) 2.32 - 1.69 = .63 units of difference between the observed and criterion abilities. (3) .63/.39 = 1.62 standard deviaitons of difference between the abilities. (4) A table of the normal distribution shows that 1 - F(1.62) = .05. Therefore, the ability value of 1.69 has a probability of .05 of coming from a normal distribution with mean = 2.32 and standard deviation = .39. Therefore, the probability that an examinee with a score of 60 has a true ability below the criterion value = .05, which is the Type II error analog that the criterion score would not be obtained by chance given the obtained ability.

Anderson et al. (1968) investigated the hypothesis that Rasch item easiness estimates are independent of the ability of the calibrating sample, and that the item easiness estimates are more stable when only items that fit the model are considered. They used the 45-item spiral omnibus intelligence test for screening applicants to the Australian Army or Royal Australian Navy. Samples of 608 recruit applicants to the Citizen Military Force (CMF) and 874 recruit applicants to the Royal Australian Navy were studied. Twelve items were deleted for zero or for 100% correct responses.

For the CMF sample, 30 items (91%) fit the model at the .01 confidence level, and 25 items (76%) fit the model at the more stringent .05 level of confidence. (The level of confidence represents the probability of obtaining the observed pattern of responses, assuming that the model is adequate to explain performance on the item.) For the Navy sample, the corresponding findings were 22 items (67%) and 16 items (48%).

The correlation between the item easiness estimates from both samples was .958 (based upon 33 items). When the items that failed to fit the model at the .05 level were deleted, the correlation increased to .990. It therefore appears that the item easiness ratios were independent of the ability of the samples from which they were computed. It should be critically noted that an intelligence test was used, and that the two subject populations probably did not differ significantly.

In a more recent study, Tinsley and Dawis (1975) gave four types of tests (verbal, numerical, picture, and item-symbol analogies) to four groups of subjects: college students, high school students, civil service clerks, and clients of the state Division of Vocational Rehabilitation (DVR). If Wright's (1967) findings could be replicated, then the ability estimates of one group should correlate highly with the ability estimates of another group for the same test. Of the 10 correlations that were computed (e.g., college students and high school students for the picture test, high school students and DVR clients on verbal analogies), all reached +.999. The invariant relationship between the ability estimates calculated for a 25-item verbal analogies test for 630 college students and 90 DVR clients replicated the relationship reported by Wright (1967) and shown in Figure 2. Tinsley and Dawis conclude that

". . . Rasch ability estimates are invariant with respect to the ability of the calibrating sample." (p. 337)

Tinsley and Dawis also investigated the degree to which the item parameters (item difficulty estimates and z-item difficulty ratios) were invariant when the analyses were performed on all items of the test. The correlation of item difficulty estimates for a given test from two examinee groups tended to be rather large (+.90). Interestingly, correlations close to zero were obtained from the DVR group with both high school and college students. This unexpected finding may be attributed to the small (n = 89) sample of DVR subjects. Generally, the item easiness ratios were invariant with respect to the ability of the calibrating sample of examinees, even though several of the comparisons used samples of questionable size.

Evaluation. The studies cited have demonstrated that if the assumptions are met, or even reasonably approximated, then person-free test calibration and item-free person measurement can be achieved by using this one-parameter logistic model. Although Hambleton and Traub (1973) report that a logistic model with an item discrimination index as a second parameter provides a better fit to their data, the inclusion of this second parameter violates true "objectivity in measurement" (Wright, 1967).

Several potential shortcomings may pose some difficulty in successfully implementing the model: (1) a pool of items must be developed that conforms to this item-analysis model, and the items must be calibrated (perhaps 20% of the items will have to be either discarded or revised); (2) the item calibration and standardization procedures require dozens of items and hundreds of subjects; (3) the model does not make direct predictions about optimal test lengths or cutting scores as do the models of Macready and Novick and Lewis; and (4) the mathematics of the model can become quite complex, posing problems for actually implementing the model and for interpretation of output. However, recent publications and the availability of computer programs (Wright & Mead, 1975, 1976) alleviate this difficulty.

The major virtues of the Rasch model can be summarized as follows: (1) Once a test has been standardized on any group of subjects, it can be given again to a different group, without the need to create parallel forms. For example, a test which had been developed by giving it to "masters" could later be given to "nonmasters." (2) All abilities will be on the same scale, regardless of the subset of items from which these abilities were estimated. Thus, person A can be measured on a hard test, and person B on an easy test.

Regression Theory

Assumptions and Rationale. The criterion-referenced testing literature has tended to emphasize the supposed dichotomy between classical test theory and the emerging CRT theory. The following discussion of regression as a means for assessing mastery is intended to point out the similarities between several CRT strategies and classical theory. Specifically, both the Bayesian and logistic models produce estimated distributions of ability, as does classical regression. A cutoff score must still be set at some point on the ability (score) distributions, regardless of what model is used to derive the distributions. This section simply portrays classical regression theory in terms of CRT theory.

The regression-theoretic approach of the "classical testing model" (Lord & Novick, 1968) describes the reason for lack of perfect mastery-nonmastery observed scores in terms of specified or estimated errors of measurement. The observed score is considered to be an unbiased estimate of an examinee's true score. It is then possible to derive a regression function that could be used to estimate true scores from observed scores. The equation for the regression function is

$$R(T|X) = r_{xx'}X + (1 - r_{xx'})m_x \ , \qquad (19)$$

where $R(T|X)$ = the true score T given the observed score X, $r_{xx'}$ = the reliability of the test, and m_x = the mean of the observed scores.

The magnitude of several types of error may also be determined. The error of measurement is the error involved when, for a randomly selected examinee, we take the observed score as an estimate of the true score. This can be expressed as $E = X - T$, and the random variable E, taking on values of e, is called the error of measurement. The standard deviation of this error of measurement, called the standard error of measurement, can be expressed in terms of the standard deviation of observed scores and the reliability of the test:

$$s_E = s_x \sqrt{(1 - r_{xx'})} \ . \qquad (20)$$

The difference between the linear regression estimate and the true score itself is called the error of estimation, and is expressed symbolically as $e = r_{xx}(x - m)(T - m_x)$. (21)

The standard deviation of these errors, called the standard error of estimation, is expressed as $s_e = s_x \sqrt{r_{xx}(1 - r_{xx})}$. (22)

Example. A graphic representation of the regression technique for a five-item test is shown in Figure 3. For each observed score, an estimated true score is obtained from $R(T|X)$, and the standard error of estimation s_e is calculated. A cutoff score based upon true scores may

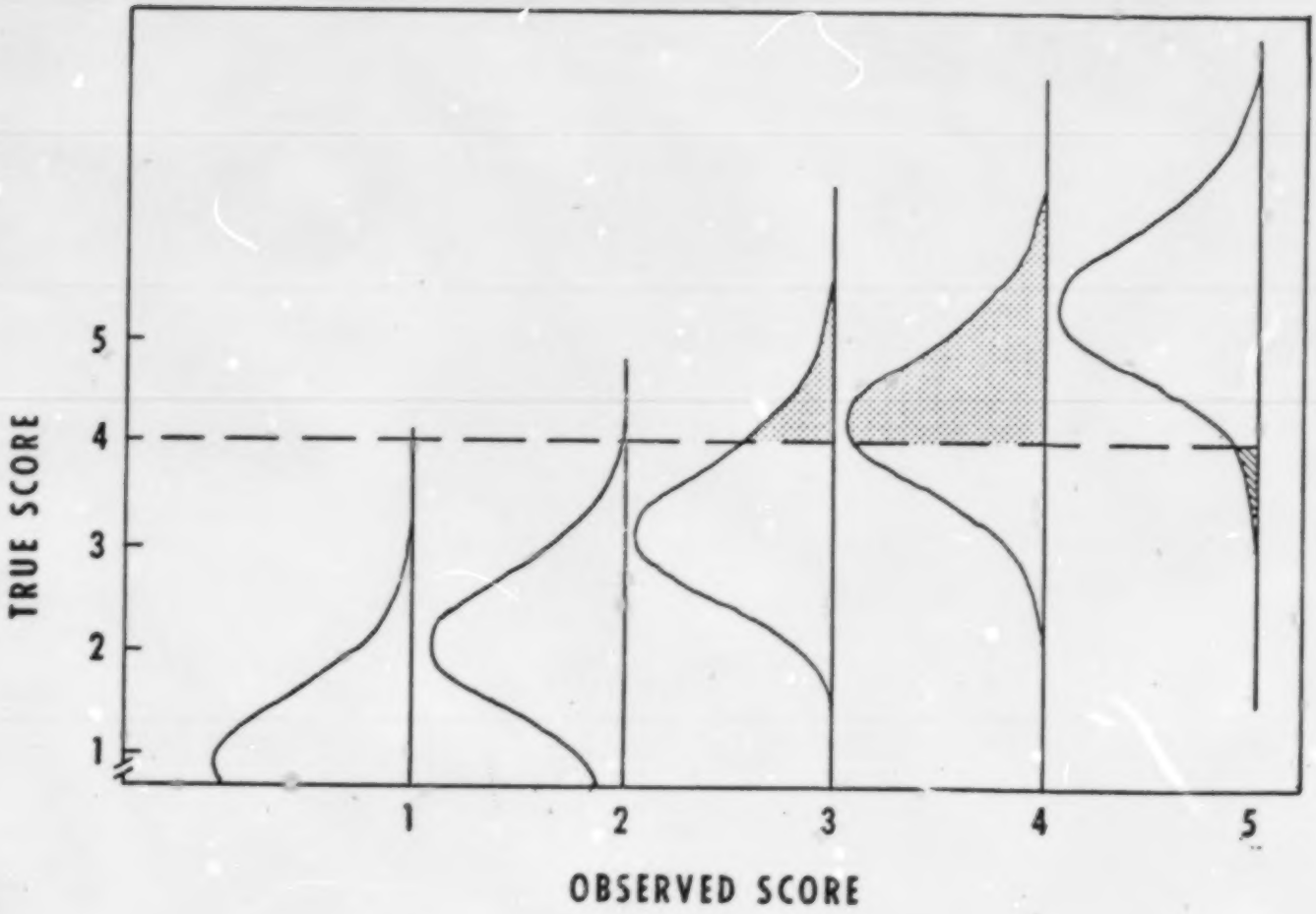


Figure 3. Regression of true score on observed score for a five-item test.

then be specified. (In this example, a true score of 4 correct has arbitrarily been chosen as the cutoff score.)

The output of the regression model, like that for the Rasch model, is a set of distributions. The mean of each distribution is the value for each $R(T|X)$, and the common standard error for all of the distributions is s_e. If the decision rule requires that all examinees be classified as masters when the value of $R(T|X)$ exceeds the criterion, and that all other scores should lead to a nonmastery decision, then the probability of misclassification can be calculated.

For persons with observed scores and estimated true scores below the criterion value, the probability that such persons might be misclassified as nonmasters is simply the proportion of the distribution exceeding the criterion value. For persons with observed scores and estimated true scores above the criterion, the probability that such persons might be misclassified as nonmasters is the proportion of the distribution below the criterion.

These probabilities of misclassification are represented as dotted and crosshatched areas, respectively, in Figure 3. If we assume that the error of estimation is normally distributed, then the probabilities can be readily obtained from a table of normal probabilities.

Two final comments are necessary. First, this procedure uses the standard error of estimate, rather than the standard error of measurement; s_e will always be smaller than s_E, since more information is used in calculating the estimated true score with a regression function than in estimating true score as the observed score. Thus, there is good reason to use the estimated true scores $R(T|X)$ in any analysis of test data. Second, the assumption of normality becomes important only when calculating misclassification errors. If the standard error of estimate cannot be assumed to be normally distributed, it may still be reported, and may prove to be useful in obtaining an estimate of the goodness of the test.

Evaluation. The regression theory approach is not a predictive model in the sense that the models developed by Dayton and Macready, Emrick, Millman, and Novick are predictive of desired test lengths and optimal cutoff scores. However, the regression approach does give probabilistic estimates of true scores, given the observed scores. The assumptions of normally distributed standard errors of estimate and of equal standard errors for all abilities may also be difficult to meet, although such departures may not pose a serious problem. And, since this is a linear regression model, it is assumed that the regression of true scores on observed scores is linear. This is a generally reasonable, though perhaps overly simplistic, assumption to make. Because the regression model has been used for many years longer than the other models reviewed in this paper, there is a greater theoretical and

empirical literature to back it up than there is for the newer, less established models. For a more technical critique of the use of regression models for estimating true scores from observed scores, see Appendix B.

SUMMARY AND CONCLUSIONS

Nature of Performance Acquisition

Performance acquisition is assumed to be an all-or-none phenomenon, according to the models developed by Emrick and by Dayton and Macready (see Table 1). Hence, these models assume that error-free test performance is also dichotomous. But the binomial, Bayesian, logistic, and classical regression models assume that performance acquisition is continuous. Performance on dichotomously scored test items must therefore be mapped onto an equivalent position on the underlying ability continuum (Roudabush, 1974). It is not possible to decide unequivocally that one assumption is more correct than the other, since the nature of performance acquisition most likely interacts with the particular type of task. Some tasks tend to elicit unitary, highly practiced, sequential behaviors, and would seem to be performed in an all-or-none fashion. Tasks which require multiskilled performances would more closely approximate the assumptions of the continuous skill acquisition models.

Measurement Error

Measurement error is defined as the difference between observed test score and true (unobservable) score that would be obtained if measurement were perfect. It is most important when one tries to infer a true "error-free" score from observed data. The Block and Crehan methods do not estimate a true score, nor do they deal directly with measurement error. Rather, they relate observed scores directly to an external criterion. Hence, any systematic error will not be a problem. But random errors which affect the consistency of observed scores will disturb the measurement process for individual cases. Fortuitously, such errors will tend to average out across groups of examinees, allowing generalizations to be made which should be valid in the "long run."

The all-or-none models deal with measurement error by stipulating values for the probability of masters committing errors and for nonmasters guessing correctly. These values are obtained by fitting the all-or-none models to observed data. Responses from both mastery and nonmastery groups can be described by binomial distributions.

The "continuous" models of Novick, Rasch, and regression theory deal with measurement error by reporting a standard error for each true

score estimate. In particular, the Rasch model provides a check on how well the model's output approximates the observed score matrix (Wright) and Mead, 1975, 1976). "Best fit" techniques are required for the Bayesian and regression models. The binomial models do not rely directly on observed data, and hence, do not deal directly with measurement error. Instead, for any hypothesized level of mastery, the models predict the observed score distribution. Adequacy of the models' predictions can be evaluated by fitting data to the hypothesized distributions. A more complete comparison of how these models are affected by measurement error must await either Monte Carlo simulation studies or considerable efforts of empirical research.

Classification Error

Unlike measurement error, classification error refers to assigning individuals to inappropriate mastery level groups--masters to the nonmastery group, and nonmasters to the mastery level group. Such errors could occur even with error-free measurement. However, measurement error interacts with classification error, further complicating the decisionmaking process of assigning examinees to mastery level groups. Suppose that, because of measurement error, all estimates of true score tended to be inflated. For a given decision rule, this would tend to decrease false negatives and increase false positives. Unfortunately, constant measurement error is the exception rather than the rule, making it virtually impossible to correct for it, and therefore separate it from classification error.

The Block and Crehan models deal with classification error empirically by comparing the decisions based on a test score with an external criterion. Hence, the classification error can be determined simply by counting the number of observed misclassifications. If examinee groups remain similar over time, these models probably provide useful and stable estimates of misclassification error.

Because none of the other models incorporates an external criterion, a direct measure of classification error is not possible. Instead, the models rely on the distributional information obtained for the estimated true scores. With this information, it is possible to predict the probability of misclassification, given various cutoff scores. Further empirical work which incorporates an external criterion is needed to verify the accuracy of such predictions.

An essential ingredient of decisionmaking on the basis of CRT scores is the concept of cost--both to the examinee and to the system which he or she is being prepared to join. Consider the case of professional licensing, such as for new medical doctors: with an extremely strict criterion, many would fail, morale would be low, and the system (society) would be deprived of much-needed medical service. However, with a very lax criterion, more examinees would pass who may not

(unfortunately) be qualified, and society would thus suffer the consequences of having "nonmasters" in practice. A similar case could be made for automobile mechanics, military medics, television repairmen, etc. Emrick's model is the only one that directly incorporates monetary costs of incorrect classifications into its procedures. However, an objective cost factor could also be incorporated into the other models quite readily. But none of the models, as developed, deals with more complex kinds of cost, such as morale, costs to society (which may have to be measured in terms of utility, not dollars), or even the cost of testing as opposed to not testing (Nader, 1976).

Test Length

For performance-oriented testing, where each item may require considerable time and expense, it is essential to be able to approximate the minimum number of items needed for good decisionmaking.

Neither the Block nor the Crehan methods explicitly deals with test length. These models were designed to show what happens when existing test results are compared to an external criterion. However, since the data are available, it would be possible to reevaluate the results, assuming that only some of the test items were used. The regression approach allows for shorter tests, but does not provide for extrapolation to longer tests.

Since the binomial model does not rely on observed data, results for tests of any length can be predicted. This aspect of the model is particularly attractive, since a first approximation to test length can be easily tried out.

The all-or-none models use observed data to help generate the necessary parameters. Once the values are available, it is possible to predict the results for tests of any length. As in the Bayesian model, such predictions will be valid only if the examinee groups remain relatively stable.

The Bayesian models can also be used as a predictor for test results of any test length. However, estimates of the values of several prior probabilities must be specified. In order for the predicted results to be applicable to real data, the estimated prior probabilities must be close approximations to the priors as determined post hoc, after data have been collected. The main feature of this model--to reduce test length as a function of increasing prior information--will be minimized to the extent that the prior information departs from correctly characterizing the population's proficiency under investigation.

The logistic model of Rasch can only be used to predict the results on a test that includes items that have already been calibrated. However, the logistic nature of the model makes it extremely powerful in

this respect. Since the item difficulty values calculated as part of the procedure are invariant across examinee groups of differing ability, any subset of items can be used with any group of examinees. Furthermore, the errors associated with each calibrated item are available, which can lead to precise predictions of classification error for tests made up of a subset of the original item pool.

Conceptualization of Mastery

The only models that explicitly define mastery are the all-or-none models. Deviations from perfection or total lack of ability are defined as measurement error. Mastery is not explicitly defined in any of the other models. Either test performance is related to some other performance (Block and Crehan) or an estimated true score on a continuum is provided. The models can then be used to evaluate test results on any specified definition of mastery.

These (continuous) models require that the tester be extremely sensitive to system requirements. If mastery is defined in terms of very high performance, then very few examinees are likely to be classified as masters; however, if mastery is defined in terms of less demanding standards, the tester (and the system) runs the risk of having a mastery group that is less than adequate. Thus, the validity of the definition of mastery in terms of the system requirements becomes a crucial issue. Empirical studies are needed in specific content areas to determine "how much ability" a master should have.

Characteristics of Items

Only the Rasch logistic model, of all the models discussed in this paper, is designed for item analysis. Other models relegate, either as assumptions or as definitions, such matters as how items are sampled, item difficulty, item homogeneity, and item independence. Certainly if an item set can be shown to violate these assumptions or definitions, the application of such a model would be questionable. Little theoretical or empirical work has been done to demonstrate the robustness of these models to violations of the assumptions.

REFERENCES

Andersen, J., Kearney, G. E., & Everett, A. V. An Evaluation of Rasch's Structural Model for Test Items. The British Journal of Mathematical and Statistical Psychology, 1968, 21, 231-238.

Beyer, W. H. (Ed.) Handbook of Tables for Probability and Statistics. Cleveland: Chemical Rubber Co., 1966.

Block, J. H. Student Learning and the Setting of Performance Standards. Educational Horizons, 1972, 183-191.

Carver, R. P. Two Dimensions of Tests: Psychometric and Edumetric. American Psychologist, 1974, 29, 512-518.

Crehan, K. D. Item Analysis for Teacher-Made Mastery Tests. Journal of Educational Measurement, 1974, 4, 255-262.

Dayton, C. M., & Macready, G. B. A Probabilistic Model for Validation of Behavioral Hierarchies. Psychometrika, 1976, 41, 189-204.

Emrick, J. A. An Evaluation Model for Mastery Learning. Journal of Educational Measurement, 1971, 8, 321-326.

Epstein, K. I., & Steinheiser, F. H. A Bayesian Method for Evaluating Trainee Proficiency. Orlando, Fla.: Proceedings of the 8th Naval Training Equipment Center/Industry Conference, November 1975.

Glaser, R., & Klaus, D. J. Proficiency Measurement: Assessing Human Performance. In R. M. Gagne (Ed.), Psychological Principles in System Development. New York: Holt, Rinehart, and Winston, 1963.

Glaser, R., & Nitko, A. J. Measurement in Learning and Instruction. In R. L. Thorndike (Ed.), Educational Measurement (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Guilford, J. P. Fundamental Statistics in Psychology and Education. New York: McGraw-Hill, 1956.

Hambleton, R. K., & Traub, R. E. Analysis of Empirical Data Using Two Logistic Latent Trait Models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 295-311.

Hershman, R. L. A Rule for the Integration of Bayesian Opinions. Human Factors, 1971, 13, 255-259.

Kaplan, M. F., & Schwartz, S. Human Judgment and Decision Processes. New York: Academic Press, 1976.

Kifer, E., & Bramble, W. The Calibration of a Criterion-Referenced Test. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

Kriewall, T. E. Applications of Information Theory and Acceptance Sampling Principles to the Management of Mathematics Instruction. University of Wisconsin Research and Development Center for Cognitive Learning: Technical Report 103, 1969.

Lord, F., & Novick, M. Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Macready, G. B., & Dayton, C. M. The Use of a Probabilistic Model in the Assessment of Mastery. Unpublished manuscript, College of Education, University of Maryland, 1975.

Meskauskas, J. A. Evaluation Models for Criterion-Referenced Testing: Views Regarding Mastery and Standard Setting. Review of Educational Research, 1976, 46, 133-158.

Millman, J. Determining Test Length: Passing Scores and Test Lengths for Objectives-Based Tests. Los Angeles: Instructional Objectives Exchange, 1972.

Millman, J. Criterion-Referenced Measurement. In W. J. Popham (Ed.), Evaluation in Education: Current Applications. Berkeley, Calif.: McCutchon Publishing Co., 1974.

Nader, R. Presentation at the 84th Annual Convention of the American Psychological Association, Washington, D.C., 1976; and reported in APA Monitor, July 1976.

Novick, M. R., & Jackson, P. H. Statistical Methods for Educational and Psychological Research. New York: McGraw-Hill, 1974.

Novick, M., & Lewis, C. Prescribing Test Length for Criterion-Referenced Measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in Criterion-Referenced Measurement. UCLA: Center for the Study of Evaluation; Monograph Series in Evaluation, 1974.

Rao, C. R. Linear Statistical Inference and Its Applications. New York: Wiley, 1965.

Rasch, G. Probabilistic Models for Some Intelligence and Attainment Tests. Copenhagen: Nielsen & Lydiche, 1960 (English translation).

Rasch, G. On General Laws and the Meaning of Measurement in Psychology. In Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics. Berkeley: University of California Press, 1961.

Rasch, G. An Item Analysis Which Takes Individual Differences into Account. British Journal of Mathematical and Statistical Psychology, 1966, 19, 49-57.

Roudabush, G. E. Models for a Beginning Theory of Criterion-Referenced Tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, 1974.

Schlaifer, R. Analysis of Decisions Under Uncertainty. New York: McGraw-Hill, 1969.

Steinheiser, F. H. Methods for Estimating Prior Probabilities. Unpublished manuscript, Army Research Institute, Alexandria, Virginia. 1976.

Tinsley, H. E., & Dawis, R. V. An Investigation of the Rasch Simple Logistic Model: Sample Free Item and Test Calibration. Educational and Psychological Measurement, 1975, 35, 325-340.

Whitely, S. E., & Dawis, R. V. The Nature of Objectivity with the Rasch Model. Journal of Educational Measurement, 1974, 11, 163-178.

Winkler, R. L. An Introduction to Bayesian Inference and Decision. New York: Holt, Rinehart, and Winston, 1972.

Wright, B. D. Sample-Free Test Calibration and Person Measurement. In Invitational Conference on Testing Problems. Princeton, New Jersey: Educational Testing Service, 1967, 85-101.

Wright, B. D., & Panchapakesan, N. A Procedure for Sample-Free Item Analysis. Educational and Psychological Measurement, 1969, 29, 23-48.

Wright, B. D., & Mead, R. J. CALFIT: Sample-Free Item Calibration with a Rasch Measurement Model. Research Memorandum No. 18, March 1975. Statistical Laboratory, Department of Education, University of Chicago.

Wright, B. D., & Mead, R. J. Rasch Model Analysis with the BICAL Computer Program. Interim Report, September 1976, to the Army Research Institute.

APPENDIX A

A GENERALIZATION OF THE EMRICK MODEL FOR THE CASE OF UNEQUAL PROPORTIONS OF MASTERS AND NONMASTERS

Kenneth I. Epstein[1]

The phi coefficient is a legitimate measure of correlation for data expressed as frequencies or proportions; it is not appropriate for conditional probabilities. The entries in the table of measurement errors proposed by Emrick and Adams (1970) and Emrick (1971a, 1971b) are conditional probabilities. A simple numerical example illustrates the type of problem which may occur if conditional probabilities are used to calculate ϕ. Assume that a group of examinees is made up of 80% masters and 20% nonmasters, that 10% of the mastery group incorrectly respond to an item, and that 5% of the nonmastery group correctly respond to the item. This situation is represented in a fourfold table in Table A-1.

Table A-1

Hypothetical Response Data for Masters and Nonmasters

True State	Observed response		
	Wrong	Correct	
Master	.10	.70	.80
Nonmaster	.15	.05	.20
	.25	.75	1.00

The phi coefficient for Table 1 is:

$$\phi = \frac{(.70)\ (.15) - (.10)\ (.05)}{\sqrt{(.80)\ (.20)\ (.25)\ (.75)}} = .5774$$

The above represents a valid use of the phi coefficient.

---

[1] My appreciation to Dr. George Macready for pointing out the problem and suggesting the direction of its solution.

We may now calculate α and β for the above data. α is defined as the probability that a nonmaster responds correctly. β is defined as the probability that a master responds incorrectly. For this example:

$\alpha = .05/.20 = .250 \qquad 1 - \alpha = .750$

$\beta = .10/.80 = .125 \qquad 1 - \beta = .875$

These data are represented in Table A-2.

Table A-2

Measurement Errors and Mastery State for Hypothetical Data

True state	Observed response		
	Wrong	Correct	
Mastery	$\beta = .125$	$1 - \beta = .5$	1
Nonmastery	$1 = \alpha = .750$	$\alpha = .250$	1
	.875	1.125	2

The phi coefficient for Table A-2 is:

$$\phi = \frac{(.875)\ (.750) - (.125)\ (.250)}{\sqrt{(1)\ (1)\ (.975)\ (1.125)}} = .6299$$

Clearly the two calculated values of ϕ are not in agreement. Table A-2 is the sort of analysis proposed by Emrick and Adams. It does not represent a valid application of the phi coefficient.

Fortunately, one can obtain a table of proportions similar to Table A-1 from a table of measurement errors similar to Table A-2, simply by multiplying each entry in the mastery row of Table A-2 by the proportion of masters, and by multiplying each entry in the nonmastery row of Table A-2 by the proportion of nonmasters. The general form for this relationship is represented in Table A-3.

Table A-3

Table of Proportions for Observed Responses and Mastery State in Terms of α, β, $P(M)$ and $P(\bar{M})$

True state	Observed response		
	Wrong	Correct	
Mastery	$P(M)\beta$	$P(M)\ (1-\beta)$	$P(M)$
Nonmastery	$P(\bar{M})\ (1-\alpha)$	$P(\bar{M})\alpha$	$P(\bar{M})$
	$P(M)\beta + P(\bar{M})(1-\alpha)$	$P(\bar{M})\alpha + P(M)(1-\beta)$	1.0

The phi coefficient for Table A-3 is derived as follows:

$$\phi = \frac{P(M)\ (1-\beta)P(\bar{M})(1-\alpha) - P(M)\beta\ P(\bar{M})\alpha}{\sqrt{[P(M)\beta + P(\bar{M})(1-\alpha)]\ [P(\bar{M})\alpha + P(M)(1-\beta)]\ P(M)P(\bar{M})}}$$

$$= \frac{P(M)P(\bar{M})\ [(1-\beta)(1-\alpha) - \beta\alpha]}{\sqrt{[P(M)\beta + P(\bar{M}) - P(\bar{M})\alpha]\ [P(\bar{M})\alpha + P(M) - P(M)\beta]\ P(M)\ P(\bar{M})}}$$

$$= \frac{P(M)P(\bar{M})\ [1-\beta-\alpha]}{\sqrt{[P(M)P(\bar{M})\alpha\beta + P(M)^2\beta - P(M)^2\beta^2 + P(\bar{M})^2\alpha + P(M)P(\bar{M}) - P(M)P(\bar{M})\beta - P(\bar{M})^2\alpha^2 - P(M)P(\bar{M})\alpha + P(M)P(\bar{M})\alpha\beta]\ P(M)P(\bar{M})}}$$

$$= \frac{P(M)P(\bar{M})\ [1-\alpha-\beta]}{\sqrt{\left[\alpha\beta + \frac{P(M)}{P(\bar{M})}\beta - \frac{P(M)}{P(\bar{M})}\beta^2 + \frac{P(\bar{M})}{P(M)}\alpha + 1 - \beta - \frac{P(\bar{M})}{P(M)}\alpha^2 - \alpha + \alpha\beta\right]\ [P(M)P(\bar{M})]^2}}$$

$$= \frac{[1-\alpha-\beta]}{\sqrt{1-\alpha-\beta+2\alpha\beta + \frac{P(M)}{P(\bar{M})}\ (\beta-\beta^2) + \frac{P(\bar{M})}{P(M)}\ [\alpha-\alpha^2]}}$$

Finally, we note that for the case where $P(M) = P(\bar{M})$, the formula above reduces to the formula given by Emrick and Adams:

$$\phi = \frac{[1 - \alpha - \beta]}{\sqrt{1 - \alpha - \beta + 2\alpha\beta + \beta - \beta^2 + \alpha - \alpha^2}}$$

$$= \frac{[1 - \alpha - \beta]}{\sqrt{1 - [\alpha^2 - 2\alpha\beta + \beta^2]}}$$

$$= \frac{[1 - \alpha - \beta]}{\sqrt{1 - (\alpha - \beta)^2}} .$$

For the example cited in the text,

$$\phi = \frac{1 - .06 - .12}{\sqrt{1 - .0036}} = \frac{.82}{.998} = .822.$$

If we have a three-item test, upon substituting into equation (3), we obtain

$$k = \frac{\log \frac{.12}{1 - .06} + \frac{1}{3}\left(\log \frac{L_2\ (.5)}{L_1\ (.5)}\right)}{\log\left(\frac{.06 \times .12}{(1 - .06)(1 - .12)}\right)}$$

$$= \frac{\log .128 + 0}{\log .0087} = .4339.$$

REFERENCES

Emrick, J. A. and Adams, E. N. An Evaluation Model for Individualized Instruction. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, Minn., 1970.

Emrick, J. A. The Experimental Validation of an Evaluation Model for Mastery Testing. Final Report: Project No. 0-A-063; Grant No. OEG-1-71-0002, U.S. Department of Health, Education, and Welfare, Office of Education, Washington, D.C., 1971.

Emrick, J. A. An Evaluation Model for Mastery Testing. Journal of Educational Measurement, 1971, 8, 321-326.

APPENDIX B

CRITIQUE OF THE SIMPLIFYING ASSUMPTIONS IN USING REGRESSION MODELS FOR ESTIMATING TRUE SCORES FROM OBSERVED SCORES

James McBride
Army Research Institute

Since $R(T|x)$ is <u>not</u> an unbiased estimator of T, the standard deviation of the error of estimate e is not the same as the conditional standard deviation of the true score for a given observed score. That is, if e is an error of estimate $(\hat{T} - T)$, then $\sigma^2(\varepsilon|x) = \sigma^2(T|x) + \text{bias}^2$. Here, $\sigma^2(T|x)$ is the conditional variance of the true scores for given observed scores, which is the distribution portrayed in Figure 3 and used for inference to the misclassification probabilities.

However, $\sigma^2(\varepsilon|x)$ (or equivalently, $\sigma^2(\varepsilon)$) is then not the appropriate variance unless there is no bias; that is, unless $E(T|\hat{T}) = \hat{T}$. And this latter relationship is generally not the case. Estimation of classification error probabilities using $\sigma^2(\varepsilon)$ as the conditional variance would therefore be inappropriate.

<u>Linear</u> regression of T on x is a convenient simplifying assumption; but in actuality, the regression may often be nonlinear. Also, the distribution of errors may seldom be normal--or even symmetrical; the same holds true for the conditional distribution of T. In sum, the estimation of error probabilities from simplified linear regression models may be considerably distorted due to the above complicating factors.

ARI Distribution List

4 OASD (M&RA)
2 HQDA (DAMI-CSZ)
1 HQDA (DAPE-PBR
1 HQDA (DAMA-AR)
1 HQDA (DAPE-HRE-PO)
1 HQDA (SGRD-ID)
1 HQDA (DAMI-DOT-C)
1 HQDA (DAPC-PMZ-A)
1 HQDA (DACH-PPZ-A)
1 HQDA (DAPE-HRE)
1 HQDA (DAPE-MPO-C)
1 HQDA (DAPE-DW)
1 HQDA (DAPE-HRL)
1 HQDA (DAPE-CPS)
1 HQDA (DAFD-MFA)
1 HQDA (DARD-ARS-P)
1 HQDA (DAPC-PAS-A)
1 HQDA (DUSA-OR)
1 HQDA (DAMO-RQR)
1 HQDA (DASG)
1 HQDA (DA10-PI)
1 Chief, Consult Div (DA-OTSG), Adelphi, MD
1 Mil Asst, Hum Res, ODDR&E, OAD (E&LS)
1 HQ USARAL, APO Seattle, ATTN: ARAGP-R
1 HQ First Army, ATTN: AFKA-OI-TI
2 HQ Fifth Army, Ft Sam Houston
1 Dir, Army Stf Studies Ofc, ATTN: OAVCSA (DSP)
1 Ofc Chief of Stf, Studies Ofc
1 DCSPER, ATTN: CPS/OCP
1 The Army Lib, Pentagon, ATTN: RSB Chief
1 The Army Lib, Pentagon, ATTN: ANRAL
1 Ofc, Asst Sect of the Army (R&D)
1 Tech Support Ofc, OJCS
1 USASA, Arlington, ATTN: IARD-T
1 USA Rsch Ofc, Durham, ATTN: Life Sciences Dir
2 USARIEM, Natick, ATTN: SGRD-UE-CA
1 USATTC, Ft Clayton, ATTN: STETC-MO-A
1 USAIMA, Ft Bragg, ATTN: ATSU-CTD-OM
1 USAIMA, Ft Bragg, ATTN: Marquat Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Lib
1 US WAC Ctr & Sch, Ft McClellan, ATTN: Tng Dir
1 USA Quartermaster Sch, Ft Lee, ATTN: ATSM-TE
1 Intelligence Material Dev Ofc, EWL, Ft Holabird
1 USA SE Signal Sch, Ft Gordon, ATTN: ATSO-EA
1 USA Chaplain Ctr & Sch, Ft Hamilton, ATTN: ATSC-TE-RD
1 USATSCH, Ft Eustis, ATTN: Educ Advisor
1 USA War College, Carlisle Barracks, ATTN: Lib
2 WRAIR, Neuropsychiatry Div
1 DLI, SDA, Monterey
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-WGC
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-MR
1 USA Concept Anal Agcy, Bethesda, ATTN: MOCA-JF
1 USA Artic Test Ctr, APO Seattle, ATTN: STEAC-MO-ASL
1 USA Artic Test Ctr, APO Seattle, ATTN: AMSTE-PL-TS
1 USA Armament Cmd, Redstone Arsenal, ATTN: ATSK-TEM
1 USA Armament Cmd, Rock Island, ATTN: AMSAR-TDC
1 FAA-NAFEC, Atlantic City, ATTN: Library
1 FAA-NAFEC, Atlantic City, ATTN: Hum Engr Br
1 FAA Aeronautical Ctr, Oklahoma City, ATTN: AAC-44D
2 USA Fld Arty Sch, Ft Sill, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: Library
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DI-E
1 USA Armor Sch, Ft Knox, ATTN: ATSB-DT-TP
1 USA Armor Sch, Ft Knox, ATTN: ATSB-CD-AD
2 HQUSACDEC, Ft Ord, ATTN: Library
1 HQUSACDEC, Ft Ord, ATTN: ATEC-EX-E-Hum Factors
2 USAEEC, Ft Benjamin Harrison, ATTN: Library
1 USAPACDC, Ft Benjamin Harrison, ATTN: ATCP-HR
1 USA Comm-Elect Sch, Ft Monmouth, ATTN: ATSN-EA
1 USAEC, Ft Monmouth, ATTN: AMSEL-CT-HDP
1 USAEC, Ft Monmouth, ATTN: AMSEL-PA-P
1 USAEC, Ft Monmouth, ATTN: AMSEL-SI-CB
1 USAEC, Ft Monmouth, ATTN: C, Facl Dev Br
1 USA Materials Sys Anal Agcy, Aberdeen, ATTN: AMXSY-P
1 Edgewood Arsenal, Aberdeen, ATTN: SAREA-BL-H
1 USA Ord Ctr & Sch, Aberdeen, ATTN: ATSL-TEM-C
2 USA Hum Engr Lab, Aberdeen, ATTN: Library/Dir
1 USA Combat Arms Tng Bd, Ft Benning, ATTN: Ad Supervisor
1 USA Infantry Hum Rsch Unit, Ft Benning, ATTN: Chief
1 USA Infantry Bd, Ft Benning, ATTN: STEBC-TE-T
1 USASMA, Ft Bliss, ATTN: ATSS-LRC
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA-CTD-ME
1 USA Air Def Sch, Ft Bliss, ATTN: Tech Lib
1 USA Air Def Bd, Ft Bliss, ATTN: FILES
1 USA Air Def Bd, Ft Bliss, ATTN: STEBD-PO
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Lib
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: ATSW-SE-L
1 USA Cmd & General Stf College, Ft Leavenworth, ATTN: Ed Advisor
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: DepCdr
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: CCS
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCASA
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACO-E
1 USA Combined Arms Cmbt Dev Act, Ft Leavenworth, ATTN: ATCACC-CI
1 USAECOM, Night Vision Lab, Ft Belvoir, ATTN: AMSEL-NV-SD
1 USA Computer Sys Cmd, Ft Belvoir, ATTN: Tech Library
1 USAMERDC, Ft Belvoir, ATTN: STSFB-DQ
1 USA Eng Sch, Ft Belvoir, ATTN: Library
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL-TD-S
1 USA Topographic Lab, Ft Belvoir, ATTN: STINFO Center
1 USA Topographic Lab, Ft Belvoir, ATTN: ETL-GSL
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATS-CTD-MS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TE
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEX-GS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTS-OR
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-DT
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-CTD-CS
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: DAS/SRD
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: ATSI-TEM
1 USA Intelligence Ctr & Sch, Ft Huachuca, ATTN: Library
1 CDR, HQ Ft Huachuca, ATTN: Tech Ref Div
2 CDR, USA Electronic Prvg Grd, ATTN: STEEP-MT-S
1 CDR, Project MASSTER, ATTN: Tech Info Center
1 Hq MASSTER, USATRADOC, LNO
1 Research Institute, HQ MASSTER, Ft Hood
1 USA Recruiting Cmd, Ft Sheridan, ATTN: USARCPM-P
1 Senior Army Adv., USAFAGOD/TAC, Elgin AF Aux Fld No. 9
1 HQ USARPAC, DCSPER, APO SF 96558, ATTN: GPPE-SE
1 Stimson Lib, Academy of Health Sciences, Ft Sam Houston
1 Marine Corps Inst., ATTN: Dean-MCI
1 HQUSMC, Commandant, ATTN: Code MTMT 51
1 HQUSMC, Commandant, ATTN: Code MPI-20
2 USCG Academy, New London, ATTN: Admission
2 USCG Academy, New London, ATTN: Library
1 USCG Training Ctr, NY, ATTN: CO
1 USCG Training Ctr, NY, ATTN: Educ Svc Ofc
1 USCG, Psychol Res Br, DC, ATTN: GP 1/62
1 HQ Mid-Range Br, MC Det, Quantico, ATTN: P&S Div

1 US Marine Corps Liaison Ofc, AMC, Alexandria, ATTN: AMCGS-F
1 USATRADOC, Ft Monroe, ATTN: ATRO-ED
6 USATRADOC, Ft Monroe, ATTN: ATPR-AD
1 USATRADOC, Ft Monroe, ATTN: ATTS-EA
1 USA Forces Cmd, Ft McPherson, ATTN: Library
2 USA Aviation Test Bd, Ft Rucker, ATTN: STEBG-PO
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Library
1 USA Agcy for Aviation Safety, Ft Rucker, ATTN: Educ Advisor
1 USA Aviation Sch, Ft Rucker, ATTN: PO Drawer O
1 HQUSA Aviation Sys Cmd, St Louis, ATTN: AMSAV-ZDR
2 USA Aviation Sys Test Act., Edwards AFB, ATTN: SAVTE-T
1 USA Air Def Sch, Ft Bliss, ATTN: ATSA TEM
1 USA Air Mobility Rsch & Dev Lab, Moffett Fld, ATTN: SAVDL-AS
1 USA Aviation Sch, Res Tng Mgt, Ft Rucker, ATTN: ATST-T-RTM
1 USA Aviation Sch, CO, Ft Rucker, ATTN: ATST-D-A
1 HQ, DARCOM, Alexandria, ATTN: AMXCD-TL
1 HQ, DARCOM, Alexandria, ATTN: CDR
1 US Military Academy, West Point, ATTN: Serials Unit
1 US Military Academy, West Point, ATTN: Ofc of Milt Ldrshp
1 US Military Academy, West Point, ATTN: MAOR
1 USA Standardization Gp, UK, FPO NY, ATTN: MASE-GC
1 Ofc of Naval Rsch, Arlington, ATTN: Code 452
3 Ofc of Naval Rsch, Arlington, ATTN: Code 458
1 Ofc of Naval Rsch, Arlington, ATTN: Code 450
1 Ofc of Naval Rsch, Arlington, ATTN: Code 441
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Acous Sch Div
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L51
1 Naval Aerospc Med Res Lab, Pensacola, ATTN: Code L5
1 Chief of NavPers, ATTN: Pers-OR
1 NAVAIRSTA, Norfolk, ATTN: Safety Ctr
1 Nav Oceanographic, DC, ATTN: Code 6251, Charts & Tech
1 Center of Naval Anal, ATTN: Doc Ctr
1 NavAirSysCom, ATTN: AIR-5313C
1 Nav BuMed, ATTN: 713
1 NavHelicopterSubSqua 2, FPO SF 96601
1 AFHRL (FT) Willsm AFB
1 AFHRL (TT) Lowry AFB
1 AFHRL (AS) WPAFB, OH
2 AFHRL (DOJZ) Brooks AFB
1 AFHRL (DOJN) Lackland AFB
1 HQUSAF (INYSD)
1 HQUSAF (DPXXA)
1 AFVTG (RD) Randolph AFB
3 AMRL (HE) WPAFB, OH
2 AF Inst of Tech, WPAFB, OH, ATTN: ENE/SL
1 ATC (XPTD) Randolph AFB
1 USAF AeroMed Lib, Brooks AFB (SUL-4), ATTN: DOC SEC
1 AFOSR (NL), Arlington
1 AF Log Cmd, McClellan AFB, ATTN: ALC/DPCRB
1 Air Force Academy, CO, ATTN: Dept of Bel Scn
5 NavPers & Dev Ctr, San Diego
2 Navy Med Neuropsychiatric Rsch Unit, San Diego
1 Nav Electronic Lab, San Diego, ATTN: Res Lab
1 Nav TrngCen, San Diego, ATTN: Code 9000-Lib
1 NavPostGraSch, Monterey, ATTN: Code 55Aa
1 NavPostGraSch, Monterey, ATTN: Code 2124
1 NavTrngEquipCtr, Orlando, ATTN: Tech Lib
1 US Dept of Labor, DC, ATTN: Manpower Admin
1 US Dept of Justice, DC, ATTN: Drug Enforce Admin
1 Nat Bur of Standards, DC, ATTN: Computer Info Section
1 Nat Clearing House for MH-Info, Rockville
1 Denver Federal Ctr, Lakewood, ATTN: BLM
12 Defense Documentation Center
4 Dir Psych, Army Hq, Russell Ofcs, Canberra
1 Scientific Advsr, Mil Bd, Army Hq, Russell Ofcs, Canberra
1 Mil and Air Attache, Austrian Embassy
1 Centre de Recherche Des Facteurs, Humaine de la Defense Nationale, Brussels
2 Canadian Joint Staff Washington
1 C/Air Staff, Royal Canadian AF, ATTN: Pers Std Anal Br
3 Chief, Canadian Def Rsch Staff, ATTN: C/CRDS(W)
4 British Def Staff, British Embassy, Washington
1 Def & Civil Inst of Enviro Medicine, Canada
1 AIR CRESS, Kensington, ATTN: Info Sys Br
1 Militaerpsykologisk Tjeneste, Copehagen
1 Military Attache, French Embassy, ATTN: Doc Sec
1 Medecin Chef, C.E.R.P.A.-Arsenal, Toulon/Naval France
1 Prin Scientific Off, Appl Hum Engr Rsch Div, Ministry of Defense, New Delhi
1 Pers Rsch Ofc Library, AKA, Israel Defense Forces
1 Ministeris van Defensie, DOOP/KL Afd Sociael Psychologische Zaken, The Hague, Netherlands

67